© 2021, American Psychological Association. This paper is not the copy of record and may not exactly replicate the final, authoritative version of the article. Please do not copy or cite without authors' permission. The final article will be available, upon publication, via its DOI: 10.1037/xge0001062

Conditionals and the Hierarchy of Causal Queries

Niels Skovgaard-Olsen

Simon Stephan

Michael R. Waldmann

Department of Psychology, Cognitive and Decision Sciences,

University of Göttingen, Germany

Author Note

Correspondence concerning this article should be addressed to Niels Skovgaard-Olsen (niels.skovgaard-olsen@psych.uni-goettingen.de, n.s.olsen@gmail.com).

Abstract

Recent studies indicate that indicative conditionals like "If people wear masks, the spread of Covid-19 will be diminished" require a probabilistic dependency between their antecedents and consequents to be acceptable (Skovgaard-Olsen et al., 2016). But it is easy to make the slip from this claim to the thesis that indicative conditionals are acceptable only if this probabilistic dependency results from a causal relation between antecedent and consequent. According to Pearl (2009), understanding a causal relation involves multiple, hierarchically organized conceptual dimensions: *prediction*, *intervention*, and *counterfactual dependence*. In a series of experiments, we test the hypothesis that these conceptual dimensions are differentially encoded in indicative and counterfactual conditionals. If this hypothesis holds, then there are limits as to how much of a causal relation is captured by indicative conditionals alone. Our results show that the acceptance of indicative and counterfactual conditionals can become dissociated. Furthermore, it is found that the acceptance of both is needed for accepting a causal relation between two co-occurring events. The implications that these findings have for the hypothesis above, and for recent debates at the intersection of the psychology of reasoning and causal judgment, are critically discussed. Our findings are consistent with viewing indicative conditionals as answering *predictive queries* requiring *evidential relevance* (even in the absence of direct causal relations). Counterfactual conditionals in contrast target *causal relevance*, specifically. Finally, we discuss the implications our results have for the yet unsolved question of how reasoners succeed in constructing causal models from verbal descriptions.

Keywords: Causality, Conditionals, Relevance, Counterfactual, Reasons

Introduction[1]

There is wide agreement that conditional statements of the type "if A, then C" play a central role in reasoning and argumentation (where 'A' refers to the antecedent and 'C' to the consequent). For instance, in 2019 much political discussion centered around the statement "If Trump is impeached, then it will affect the 2020 election". At the same time, conditionals pose many unsolved theoretical problems that have kept researchers busy, despite continuous, multidisciplinary efforts (Bennett, 2003; Kern-Isberner, 2001; Kratzer, 2012; Nickerson, 2015; Oaksford & Chater, 2010a; Spohn, 2013).

One of the reasons why conditionals are thought to be so central in our cognitive lives is due to their relationship with causal knowledge (Oaksford & Chater, 2010b). The linguistic encoding of knowledge about causal relations in conditionals plays a vital role for the cultural transfer of causal knowledge across generations. For causal knowledge about objects that are not in our immediate vicinity, we rely on culturally transferred causal knowledge. The same goes for objects that are governed by mechanisms, which we do not fully understand, like artifacts designed by engineers. In addition, the acquisition of causal knowledge through observed covariances and interventions dealing with the objects that *are* in our direct vicinity is often guided by linguistically acquired causal schemes (Gopnik et al., 2004). Various authors have emphasized that probably most of our causal knowledge comes through this linguistic source (e.g. Pearl, 2009, Ch. 7). But according to Danks (2014, Ch. 4), it is also the one that is the least investigated empirically.

The relationship between conditionals and causal relations has, however, been the focus of much theoretical discussion. The importance of this issue is highlighted by counterfactual approaches to causation coming from philosophy (Goodman, 1947; Lewis,

[1] We would like to thank Dominik Glandorf, Louisa Reins, and Maike Holland-Letz for their help in coding responses and setting up experiments. We also thank audiences at talks at London Reasoning Workshop (2019), EuroCogSci (2019), Regensburg University (2020), and the Reviewers and our Editor, Pierre Barrouillet, for valuable feedback.

1973; Collins, Hall, & Paul 2004), computer science (Pearl, 2009), and statistics (Morgan & Winship, 2018; VanderWeele, 2015). Recently, various authors in psychology and philosophy have also made a case for causal interpretations of indicative conditionals (e.g. Oaksford & Chater, 2017; Andreas & Günther, 2018; van Rooij & Schulz, 2019; Vandenburgh, 2020).

In this paper, we investigate whether indicative conditionals by themselves suffice to express causal relations or whether there are aspects of causal relations that are not captured by indicatives.[2] We will rely on Pearl's (2009) theory of causality and his idea of *a hierarchy of causal queries*. Through our experiments, we present new evidence in support of this framework and investigate its relations to natural language conditionals. Before we turn to our research questions, we first sketch some recent developments in the psychology of reasoning, which have kindled a renewed debate about the causal interpretation of indicative conditionals. Secondly, we outline Pearl's theory of a hierarchy of causal queries and discuss its critical potential vis-à-vis this debate.

Indicative Conditionals and Probabilities

Building on the work of Adams (1975), Edgington (1995), and Bennett (2003), psychologists have found support for the hypothesis that:

$$[\text{Eq1.}] \qquad P(\text{if A, then C}) = P(C|A),$$

which goes by the name of "the Equation" or "the conditional probability hypothesis" (Evans, Handley, & Over, 2003; Oberauer & Wilhelm, 2003; Over, Hadjichristidis, Evans, Handley, & Sloman, 2007; Pfeifer & Kleiter, 2009). Recently, these results were challenged, however.

[2] As a short-form, we refer to indicative conditionals, like "If A, then C", as 'indicatives', and to counterfactual conditionals, like "If A had not been the case, then C would not have been the case", as 'counterfactuals'. Our focus will be on paradigmatic cases of indicative conditionals, like the examples provided in the main text. Other controversial examples like non-interference conditionals ("If Trump won the 2020-election, then pigs can fly!") are not treated here but see Douven (2016) and Skovgaard-Olsen (2016) for further discussion.

It has been found that the relationship between P(if A, then C) and P(C|A) is moderated by *relevance effects* of the probabilistic dependency between A and C (Skovgaard-Olsen, Collins, et al., 2019; Skovgaard-Olsen, Kellen, et al., 2017; Skovgaard-Olsen, Singmann, & Klauer, 2016; Vidal & Baratgin, 2017). This type of probabilistic dependency can be captured by ΔP as a measure of the extent to which A changes the probability of C:

$$[\text{Eq2.}] \qquad \Delta P = P(C|A) - P(C|\neg A)$$

These studies have found that in the case of Positive Relevance, ($\Delta P > 0$), the conditional probability remained a good predictor of both the acceptance and probability of indicative conditionals. An example would be "If Paul pushes down the gas pedal, then the car will speed up" in the context of a scenario describing Paul driving in his car and running late for work. For cases of Negative Relevance ($\Delta P < 0$) and Irrelevance ($\Delta P = 0$) this relationship was disrupted, however. Two examples would be "If Paul pushes down the gas pedal, then the car will slow down" (Negative Relevance) and "If Paul is wearing a shirt, then his car will suddenly break down" (Irrelevance).

These findings suggest that participants tend to view indicative conditionals as defective if their antecedents fail to raise the probability of their consequents. In such cases, their antecedents fail to provide a reason *for* the consequent (Douven, 2016; Krzyżanowska, Collins, et al., 2017; Skovgaard-Olsen, 2016; Spohn, 2013). Drawing on the literature on confirmation measures, the notion of A being a reason *for* or *against* C is here explicated in terms of its *evidential relevance*, or the difference in degrees of beliefs that A makes to C (Spohn 2012, Ch. 6). If A raises the probability of C ($\Delta P > 0$), then A is said to be a reason *for* C, or *positively relevant* to C. If A lowers the probability of C ($\Delta P < 0$), then A is said to be a reason *against* C, or *negatively relevant* to C. If A leaves the probability of C unchanged ($\Delta P = 0$), then A is said to be *irrelevant* to C, or neither a reason *for* nor *against* C. Indicative conditionals are said to express such qualitative reason relation assessments on this account (Brandom, 1994; Spohn, 2013; Skovgaard-Olsen, 2016; see also Rott, 1986; Krzyżanowska,

Wenmackers, et al., 2013; Douven, 2016). Throughout the paper, we will measure qualitative assessments of the extent to which A is a reason for/against C on an ordinal scale and refer to them as 'ordinal reason relation assessments'.

As a psychological construct, it is possible that multiple factors influence the assessment of relevance and reason relations including topical relevance, processing effort, and goals in a dialogue (Walton, 2004; Wilson & Sperber, 2004). Potentially, such factors influence the categorization of variables as *capable* or *incapable* of affecting the probability of the consequent. Variables that are categorized as *incapable* get ignored. This makes it seem defective to find such variables in the antecedent of conditionals, where one expects to find a *reason for* the consequent (Skovgaard-Olsen, Collins et al., 2019). As a measure of the cognitive effects of a variable, we rely on the notion of probabilistic difference-making from above but note that there is a discussion with mixed evidence concerning further factors influencing the perceived relevance.[3]

The data pattern described above constitutes the Relevance Effect as an interaction effect (see Figure 1).

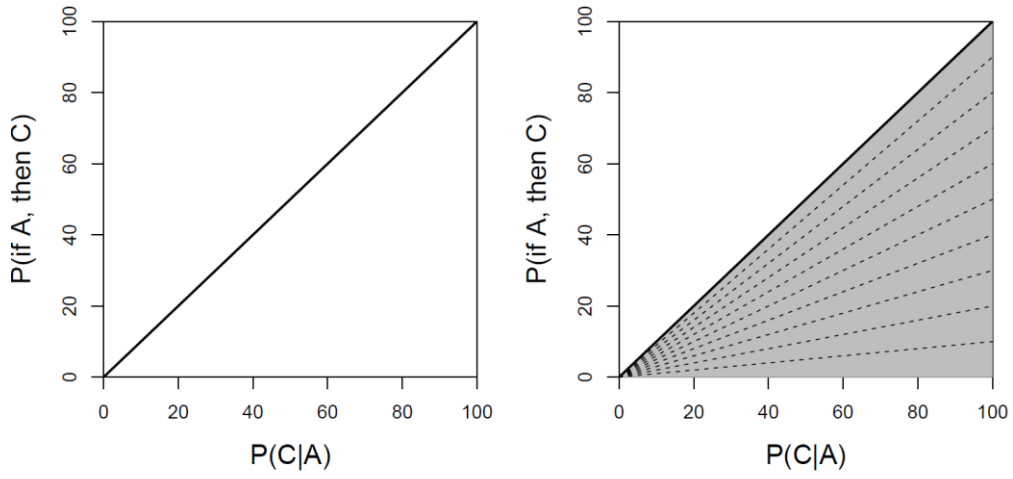


Figure 1. The left panel illustrates relationship predicted by [Eq1.]. The right panel illustrates the Relevance Effect, i.e. the moderation of the slope by relevance, in case of irrelevance ($\Delta P = 0$) or negative relevance ($\Delta P < 0$), after Skovgaard-Olsen, Kellen et al. (2019).

[3] See e.g. Cruz et al. (2016), Skovgaard-Olsen, Singmann et al. (2017, supplementary materials), Vidal and Baratgin (2017), Krzyżanowska et al. (2017).

Accounts differ on whether this finding is to be given a semantic or pragmatic interpretation (see e.g. Skovgaard-Olsen, Collins, et al., 2019 for a review), but here we focus on a different issue. It has recently been suggested (e.g. in Oasksford & Chater, 2020a, 2020b; van Rooij & Schulz, 2019) that relevance effects of this kind need to be given a causal interpretation. One of the goals of the present paper is to systematically explore this link through a series of experiments.

As we will explain further below, these experiments have a bearing on whether (1) P(C|A) is a good predictor of P(if A, then C) as predicted by [Eq1.] (Evans & Over, 2004; Oaksford & Chater, 2017), (2) whether a causal interpretation (van Rooij & Schulz, 2019; Oaksford & Chater, 2020a, 2020b) or (3) an evidential relevance interpretation of P(if A, then C) is needed (Skovgaard-Olsen, Singmann, & Klauer, 2016). According to Evans and Over (2004), people assess P(C|A) via the Ramsey Test:

> RAMSEY TEST: to evaluate 'if A, then C' add the antecedent (i.e. A) to the set of background beliefs, make minimal adjustments to secure consistency, and evaluate the consequent (i.e. C) on the basis of this temporarily augmented set.

Using the Ramsey Test as a basis of explicating the relationship between conditionals and suppositional reasoning has been influential in at least three competing research programs in logic (Horacio, 2007). However, in and of itself it is an abstract description of a mental algorithm which needs to be fleshed out in terms of psychological processes to be of use for cognitive scientists. As Over et al. (2007) have noted:

> Explaining how the Ramsey Test is actually implemented—by means of deduction, induction, heuristics, causal models, and other processes—is a major challenge, in our view, in the psychology of reasoning. (p. 63)

In the past decade, psychologists have made extensive use of the Ramsey Test (for a review, see Oaksford & Chater, 2020a). But the fundamental problem that Over et al. (2007) pointed to remains. Resolving this issue is important, because [Eq1.] and the abovementioned

probabilistic view on conditionals has not just been taken to be one view on conditional reasoning among others. Rather, it has been treated as "one of the defining features of what has come to be referred to as the *new paradigm* in cognitive psychology" (Nickerson, 2015, p. 199) and been said to be "at the heart of the probabilistic *new paradigm* in reasoning" (Oaksford & Chater, 2017, p. 330; see also Vance & Oaksford, 2020).

One of the processes for implementing the Ramsey Test that Over et al. (2007) consider is the use of *causal models*. In line with this, Fernbach, Darlow, et al. (2011) and others have argued that causal beliefs are used as a guide for estimating subjective probabilities. The notion that conditional probabilities are assessed based on causal models via the Ramsey Test is interesting. If it can be corroborated, then this would have implications for which of the previously mentioned interpretations relating P(C|A) and P(if A, then C) is correct. For if the conditional probabilities estimated via the Ramsey Test were to rely on causal models, then P(C|A) would not be independent of a causal interpretation. In that case, P(if A, then C) would also not be independent of causal considerations given [Eq1.].

In addition, recent work on causal power suggests another possible connection between indicative conditionals and causality, which we will now turn to, because it will figure centrally in our later experiments.

Causal Power and Alternative Causes

On Cheng's (1997) account of causal power, the generative power of a cause to produce its effect is explicated by a scaled version of ΔP, where the causal contribution of alternative causes is shielded off:

[Eq3.] $$W_{Cause} = \frac{\Delta P}{1-P(effect|\neg cause)}\ ,\ \Delta P = P(effect|cause) - P(effect|\neg cause)$$

Causal power (W_{Cause}) is here understood as the probability with which a target cause

generates its effect[4] independently of alternative causes: P(effect|cause,¬alternatives). [Eq3.] measures this quantity by determining how much the candidate cause contributes to raising the probability of the effect, while bracketing the influence of alternative causes. Following Glymour (2001), causal power has been used to parameterize Bayes nets (see e.g. Griffiths & Tenenbaum, 2005; Fernbach, Darlow, et al., 2010, 2011; Fernbach & Erb 2013; Cummins, 2014; Meder, Mayrhofer, et al., 2014; Aßfalg & Klauer, 2019; Stephan & Waldmann, 2018), as illustrated in Figure 2:

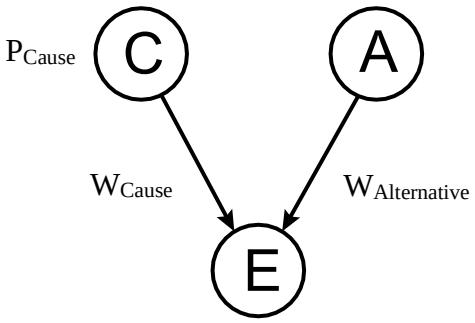


Figure 2. Common-effects Bayes Net, parameterized by the base-rate (P_{Cause}) of the cause, C, its causal power (W_{Cause}), and the combined base-rate and causal power ($W_{Alternatives}$) of the alternative cause(s), A. 'E' = effect.

Here 'C' refers to the cause and 'A' refers to alternative causes. Throughout this paper, we follow, however, the convention of using 'A' and 'C' to refer to the antecedent and consequent of conditionals, whether or not they are related as cause and effect. Based on this parametrization and other assumptions (discussed in Luhmann & Ahn, 2005), conditional probabilities have been explicated as follows, with 'W' representing the causal powers of the respective causes:

$$[Eq4.] \qquad P(effect|cause) = W_{cause} + W_{alternative} - W_{cause} * W_{alternative}$$

Notice how conditional probabilities are here explicated in terms of causal power parameters, which in turn are defined via conditional probabilities. There is accordingly a choice as to which of these constructs (i.e. conditional, subjective degrees of belief or mental

[4] For preventive causes, a separate equation was given by Cheng (1997), which we return to in Experiment 1 (see [Eq5.]).

representations of causal powers) is to be treated as psychologically primitive. For example, for Cheng (1997) causal powers represent latent, causal capacities of distal objects. On this view, the relative frequencies encoded in conditional probabilities are merely the manifestations of these latent capacities. But this is not the only position possible and the answer to the question of psychological primacy will have repercussions for the relationships between conditionals, conditional probabilities, and causality.

Oaksford and Chater (2017) have suggested that a causal interpretation of indicative conditionals can be combined with work in probabilistic treatments of conditionals based on the Ramsey Test (e.g. Adams, 1975; Edgington, 1995; Bennett, 2003; Evans & Over, 2004; Oaksford & Chater, 2007). Oaksford and Chater (2017) do this by combining the thesis P(if A, then C) = P(C|A) [Eq1.] with a causal power explication of conditional probabilities (see [Eq4.]). Making this move allows Oaksford and Chater (2017) to emphasize that there is an inferential dependency between antecedents and consequents of indicative conditionals (in line with, e.g., Douven, 2016; Krzyżanowska, Collins, et al., 2017; Skovgaard-Olsen, Singmann, et al., 2016; Spohn, 2013). At the same time, it allows Oaksford and Chater (2017) to build on the work on probability logic of Adams (1975), which has been applied to the psychology of reasoning (e.g. in Evans & Over 2004; Oaksford & Chater, 2007; Pfeifer & Kleiter, 2009).

One challenge to this account, however, is that the Relevance Effect (Skovgaard-Olsen, Singmann, et al., 2016) identifies boundary conditions on P(C|A) as a predictor of P(if A, then C). As a consequence, if probabilistic dependency is factored into the account through a causal power explication of conditional probabilities, then we are left without an account of why relevance moderates the relationship between P(C|A) and P(if A, then C) in violation of [Eq1.]. The interaction effect depicted in Figure 1 shows that P(if A, then C) can vary due to the influence of relevance even when P(C|A) is held constant.

Accordingly, Oaksford and Chater (2020b) discuss the different possibility where the

Relevance Effect is itself an indicator of a causal interpretation of indicative conditionals. But this amounts to abandoning [Eq1.] in its full generality.

As noted by van Rooij and Schulz (2019), there is, however, also a different possibility for interpreting the relationship between conditional probabilities, causal power, and P(if A, then C). The general account relies on interpreting the acceptability of indicative conditionals in terms of causal power. But by introducing this conjecture, van Rooij and Schulz rely on the auxiliary hypothesis that participants tend to ignore alternative causes. The motivation for this auxiliary hypothesis is that the equation for causal power [Eq3.] shows that causal power coincides with the conditional probability of the effect given the cause when there are no alternative causes:

$$\{\text{x: x is an alternative cause of } E\} = \emptyset \implies W_{\text{Cause}} = P(\text{effect}|\text{cause})$$

If participants ignore alternative causes and by mistake treat P(effect|¬cause) as 0, then they should also underestimate P(effect|cause) by evaluating it as P(effect|cause, ¬alternatives). Their estimate of P(effect|cause) will then coincide with the value of causal power, which would explain the studies corroborating [Eq1.]. In van Rooij and Schulz (2019), an formal analysis of such limiting cases was used to propose a causal power measure of the acceptability of conditionals by arguing that it is the presence of causal power that makes indicative conditionals acceptable.

Studies in the psychology of causal judgments have shown that reasoners often tend to neglect alternative causes (see, e.g. Rottman & Hastie, 2014, for an overview). These findings, in turn, fit with well-known effects from the psychology of reasoning concerning inferences like denial of the antecedent (*If A, C; ¬A, therefore ¬C*) and affirmation of the consequent (*If A, C; C, therefore A*). Indeed, a neglect of alternative antecedents (e.g. "If B, then C") has long been suspected as being part of the explanation why participants would endorse these logically fallacious inferences (Cummins, 1995; Politzer & Bonnefon, 2006). In linguistics, there is a convergent body of research studying conditional perfection (for review,

see Liu, 2019), which describes the tendency to strengthen an indicative conditional into a bi-conditional that suppresses alternative antecedents. Moreover, the tendency to suppress the impact of alternative hypotheses has long been suspected of playing a role in the confirmation bias (Nickerson, 1998).

According to Fernbach, Darlow, et al. (2010, 2011), participants who are asked for conditional probabilities report them but are biased by their neglect of alternative causes. Alternatively, one may hold that participants who are asked for conditional probabilities construe the task differently and give causal power estimates instead (but see Aßfalg & Klauer, 2019). For our purposes, it is, however, interesting to note that if participants tend to ignore alternative causes, then the causal power interpretation of indicative conditionals in van Rooij and Schulz (2019) can be used to account for the Relevance Effect.

Accordingly, van Rooij and Schulz conjecture that what explains when P(C|A) is and when it is not a good predictor of P(if A, then C) in studies like Skovgaard-Olsen, Singmann, et al. (2016) is exactly whether participants take alternative causes into account. Participants are thereby portrayed as ignoring alternative causes when processing positive relevance conditionals, like "If Paul pushes down the gas pedal, then the car will speed up". In contrast, participants are predicted to take alternative causes into account when processing irrelevance items, like "If Paul is wearing a shirt, then his car will function normally", where the antecedent is obviously not an appropriate cause.

In Experiment 1, we test whether participants' tendency to ignore alternative causes makes them estimate P(C|A) as causal power in scenarios that can be interpreted causally. Experiment 1 thereby provides a critical test of the following hypotheses based on van Rooij and Schulz's (2019) work:

(H_1) causal power [Eq. 3] accounts for the acceptance of indicative conditionals.

(H_2) participants' tendency to ignore alternative causes is part of the explanation of the Relevance Effect.

We now turn to Pearl's (2009) theory of causality, which we will use to reconceptualize the relationship between indicative conditionals and causal relations. Of central importance in this context is the following observation. While Oaksford and Chater (2017) and van Rooij and Schulz (2019) argue for a causal interpretation of indicative conditionals, Pearl's idea of a hierarchy of causal queries invites a more complex picture in which indicative conditionals only play a partial role.

Pearl's Hierarchical Theory of Causality

According to Pearl (2009) and Pearl and Mackenzie (2018), there are three conceptual layers of causality: *prediction*, *intervention*, and *counterfactual dependency*. An understanding of these three conceptual layers is manifested by the ability to answer three different types of queries concerning the relationship between two variables, X and Y. In Pearl and Mackenzie (2018), these queries take, roughly, the following form:

Table 1. The Hierarchy of Causal Queries

Query Type	Natural Language Query	Computational Model
Predictive	"What happens to my belief in Y if I see X?"	Bayes net, SEM
Interventional	"What happens to Y if I do X?"	causal Bayes net, SEM
Counterfactual	"Would Y not have occurred if X had not occurred?"	SEM

Note. SEM = Structural Equation Modelling (see Appendix A). The distinction between 'Bayes nets' and 'causal Bayes nets' is made to emphasize that Bayes nets exist with both undirected edges representing symmetrical relations of evidential relevance, as well Bayes nets that encode directed edges used for representing assymmetrical relations of causal relevance (Højsgaard, Edwards, et al., 2012; Danks, 2014).

In Pearl (2009, p. 29), the following examples are given: 1) "would the pavement be slippery if we *find* the sprinkler off" (*prediction*), 2) "would the pavement be slippery if we *make sure* that the sprinkler is off" (*intervention*), and 3) "would the pavement be slippery *had* the sprinkler been off, given that the pavement is in fact not slippery and the sprinkler is on?" (*counterfactual*). As a normative competence model of causal inference, Pearl (2009) presents a theory of causal Bayes nets augmented by structural equation modelling (SEM). For Pearl (2009), it is important to emphasize that there are three irreducible layers of conceptual understanding of causal relations: 1) statistical associations for predictive inference (which

can be computed by conditionalization, e.g. via Bayes nets), 2) predictions based on interventions (which are observed through manipulations in randomized, experimental studies),[5] and 3) counterfactual inferences (which can only be computed based on structural equation models of the data generating processes). In Appendix A, we illustrate the distinction between these computational models via one of Pearl's examples.

Several aspects of Pearl's theory have been investigated in psychological studies. For instance, whether reasoners differentiate between observational probabilities and interventional probabilities (Sloman & Lagnado, 2005; Waldmann & Hagmayer, 2005). Similarly, studies have looked at participants' understanding of the Markov assumption and the implied conditional independencies (Rehder, 2014; Rottman & Hastie, 2014; Mayrhofer & Waldmann, 2015). But whereas the causal Bayes net component of the theory has received extensive attention, the structural equation component has received less attention in psychology. Yet, some exceptions like Lagnado, Gerstenberg, et al. (2014) do exist. In Appendix A, we explain why it is important for psychology to focus more on SEM.

Research Questions Motivating this Investigation

The central question motivating the present inquiry is this: what role do conditionals as linguistic expressions play in representing causal information? Or: by accepting a conditional statement in a causal scenario, which of the three aspects of the causal relation highlighted by Pearl does a reasoner thereby accept, if any? Looking back at Table 1, answering the first two types of queries seems[6] equivalent to processing indicatives ("will the pavement be slippery,

[5] In addition, these interventions can now also be computed by applying Pearl's (2009) do-calculus to observational studies (see also Morgan & Winship, 2018).

[6] Note that Pearl (2009, p. 29) uses "would" instead of "will" in the consequents of the observational and interventionist queries. However, the resulting conditional questions are closer in meaning to the indicatives above given the indicative antecedents than the corresponding counterfactuals. When Pearl wants to stress a counterfactual interpretation, he often uses "would have" (see e.g. Pearl & Mackenzie, 2018, p. 320).

if we see/make sure that the sprinkler is off?"). Moreover, answering the third type of query is naturally taken to involve processing counterfactuals ("would the pavement have been slippery, if the sprinkler had been off?"). It is then natural to formulate the following hypothesis based on Pearl's view:

(H_3) causal relations encode multiple layers, some of which *can* be expressed by indicatives (i.e. predictive queries), whereas the most advanced one requires the use of counterfactuals (i.e. counterfactual queries).

This, in turn, makes it natural to conjecture that:

(H_4) indicatives that *support* and indicatives that *do not support* counterfactuals can be empirically distinguished (see also Lassiter, 2017).

(H_5) the use of indicatives and the acceptance of causal relations can be dissociated even in causal scenarios.

To illustrate (H_5), indicatives based on spurious correlations can be used to answer predictive queries, but they do not express direct causal links between their antecedents and consequents. A well-known example is "If the barometer falls, then bad weather is coming". According to (H_4), we would expect that a characteristic of such indicative conditionals expressing spurious correlations is that they do not support counterfactuals.

Depending on the query, the intervention might represent a natural continuation expressed in the indicative mood (e.g. "the cappuccino will taste better, if I use espresso beans"). Alternatively, the intervention might represent an unlikely continuation expressed in the subjunctive mood (e.g. "the cappuccino would taste better, if I bought an espresso machine for 10.000 €"). In our experiments, we are less concerned with interventions, however. Instead, we focus instead on different aspects of the distinction between predictive use of indicative conditionals for expressing statistical associations of *evidential relevance* and use of counterfactuals to answer queries that target *causal relevance*. For a psychological theory of probabilistic reasoning, ΔP is often used to represent evidential relevance and causal

power can be used to represent causal relevance.

Overview of the Experiments

To address the above research questions, we conducted experiments that contrast a situation in which participants are provided a detailed representation of a mechanism linking inputs and outputs with observations of blackbox trials in which the mechanism was covered. The animations were inspired by the 1993 computer game, "The Incredible Machine". Illustrations of the trials are shown in Figures 3 and 4 below:

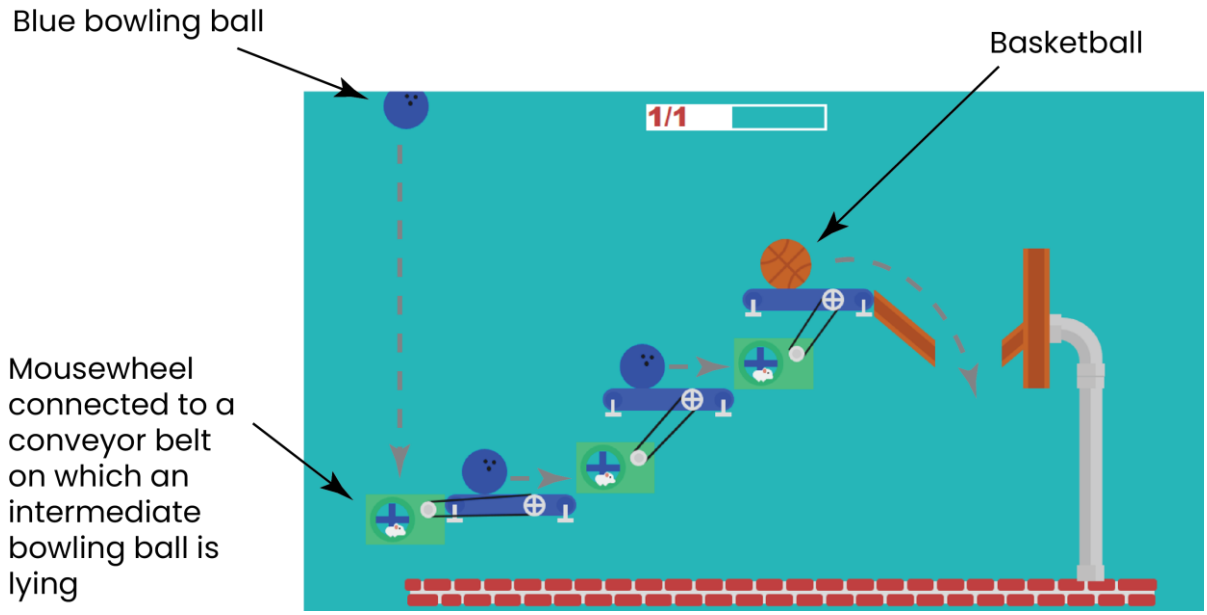


Figure 3. Annotated illustration of a Machine Trial in which the whole mechanism is visible. See https://osf.io/fa9rj/ for a video illustration.

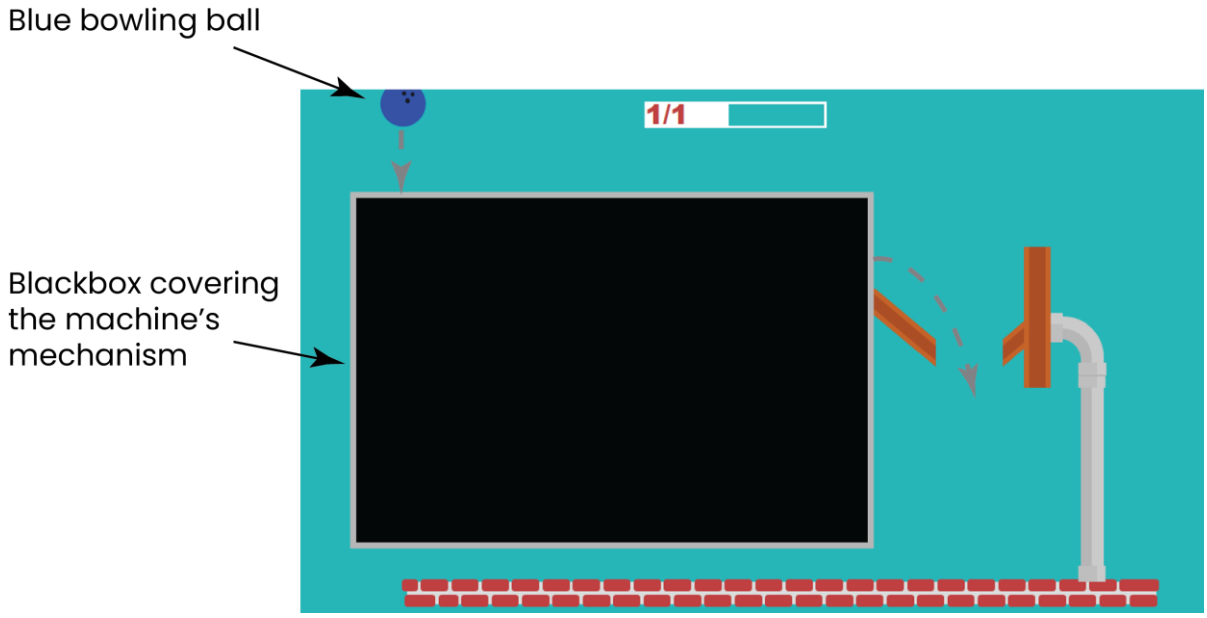


Figure 4. Annotated Illustration of a Blackbox Trial in which the mechanism is covered. See https://osf.io/fa9rj/ for a video illustration.

Figures 3 and 4 show annotated snapshots of the animations. Figure 3 depicts the Machine condition in which a causal chain unfolds when a blue bowling ball (root cause) falls onto a mouse wheel connected to a conveyor belt. This chain of events ends with the basketball dropping into the basket. In Figure 4, the mechanism is concealed. Note that this system is not deterministic because the mice can start to run on their own and they may sometimes not run even if a bowling ball hits their cage. We adopted this format as a way of manipulating the depth of participants' understanding of a causal relation in light of long-standing debates in the psychology of causal judgment about possession of structural knowledge that goes beyond associative learning (Waldmann, 1996; Waldmann & Hagmayer, 2005; Pelley, Griffiths, et al., 2017).

The animations that we used conveyed the information in a trial-by-trial format. Usually, the psychology of reasoning (Manktelow, 2012) follows the research tradition on cognitive illusions (Kahneman, Slovic, et al., 1982) in studying reasoning problems via verbal scenarios. However, trial-by-trial learning paradigms are common in areas such as the psychology of learning (Bouton, 2016) and causal reasoning (Waldmann, 2017). The finding of the description-experience gap (Hertwig & Erev, 2009; Rehder & Waldmann, 2017) shows that the two paradigms can lead to different results. There is therefore a need for applying trial-by-trial learning paradigms to problems in the psychology of reasoning (Vance & Oaksford, 2020).

In our experiments, we manipulated different levels of contingency (ΔP), conditional probability ($P(C|A)$), and causal power (W_{Cause}). A trial-by-trial learning paradigm with the animated mouse-wheel machine was used in Experiments 2-6. Table 2 provides a brief overview of the experiments:

Table 2. Overview of the Experiments

Exp	Purpose	Method	Hypothesis
1	Critical test of assumptions needed to account for the Relevance Effect based on van Rooij and Schulz (2019).	Verbal scenarios, test of causal power as a predictor of P(if A, then C) and the influence of alternative causes on the Relevance Effect.	H_1, H_2
2	Replication of the Relevance Effect in a trial-by-trial learning paradigm.	Animations with the mouse-wheel machine task in a causal chain structure.	See below.
3	Investigate the relationship between judgments of causal power, indicatives, counterfactuals, and singular causation.	Animations with the mouse-wheel machine task in a causal chain structure with a blackbox condition.	H_3
4	Test of the acceptance of indicatives and counterfactuals as predictors of singular causation judgments.	"	H_3
5	Test of dissociation between the acceptance of indicatives and counterfactuals.	Animations with the mouse-wheel machine task in a common cause structure with a blackbox condition.	H_4, H_5
6	Replicating Experiment 4 while controlling for the influence of tense and the order of events.	"	H_4, H_5

Using the verbal stimulus materials used to originally document the Relevance Effect in Skovgaard-Olsen, Singmann, et al. (2016), Experiment 1 aimed at providing a critical test of assumptions in van Rooij and Schulz (2019). Experiment 1 thereby probed a causal power account of the acceptance of indicative conditionals (H_1) and whether participants' tendency to ignore alternative causes accounts for the Relevance Effect (H_2).

The goal of Experiment 2 was to test whether the Relevance Effect could be replicated in a trial-by-trial learning task.

The next two experiments involved singular causation judgments. Singular causation judgments typically concern situations in which both the potential cause and effect are known to have co-occurred and reasoners have to establish whether the former actually caused the effect on this specific occasion. Our interest in these types of judgments originates in their role in testing (H_3) – with its claim of multiple conceptual layers in the understanding of causal relations. Moreover, we investigated singular causation judgments to ensure that participants were making the causal attributions intended by our experimental designs.

Experiment 3 investigated whether the four central constructs of 1) causal power, 2)

indicative conditionals, 3) counterfactual conditionals, and 4) singular causation are influenced by the same factors in a large between-subjects experiment. The motivation for this comparison was that according to a causal interpretation of conditionals, one would expect conditionals to be affected by manipulations that influence causal judgments.

The purpose of Experiment 4 was to investigate whether singular causation judgments could be predicted by the acceptance of indicative and counterfactual conditionals. In line with the hierarchy of causal queries, Pearl (2009, Ch. 10) and Halpern (2019) build in explicit counterfactual conditions in their accounts of singular causation. Experiment 4 therefore tests whether the acceptance of counterfactual conditionals plays a role for singular causation.

Experiments 5 and 6 compared the acceptance of indicative and counterfactual conditionals in a common-cause version of the trial-by-trial learning paradigm. The goal was to investigate whether the acceptance of indicatives and counterfactuals would become dissociated for diagnostic and common-cause conditionals to test (H_4) and (H_5). The investigation of common-cause and diagnostic reasoning scenarios is crucial because they exemplify cases, where the answers to predictive queries need not represent relations of direct causal impact. For instance, measurements on a barometer are diagnostic for the coming weather conditions and can be used to answer predictive queries (e.g. "Can we expect bad weather, if the barometer falls?"). But the common cause of both are changes in atmospheric pressure.

Experiment 1

According to van Rooij and Schulz (2019), the acceptability of indicative conditionals is determined by causal power (H_1). Based on this account, it is natural to conjecture that participants assign probabilities to indicative conditionals, 'if A, then C', based on causal power.[7] On the auxiliary assumption that participants ignore alternative causes, causal power

[7] Note that van Rooij and Schulz (2019) are careful in stating their theory only in terms of categorical acceptance of indicative conditionals. But they indicate an extension of it to

would coincide with the conditional probability, as we have seen. van Rooij and Schulz (2019) suggest (H_2) that we can use this observation to account for the Relevance Effect in Skovgaard-Olsen et al. (2016). To do so, one would have to conjecture that participants' tendency to ignore alternative causes makes P(C|A) a good predictor of P(if A, then C) for Positive Relevance ($\Delta p > 0$) items. In contrast, the lack of causal dependence of consequent on the antecedent would make P(C|A) overestimate P(if A, then C) for Irrelevance items ($\Delta p = 0$). In addition, we probe whether we can replicate the Relevance Effect in a situation, where it is difficult to ignore alternative causes by using a task that builds on Byrne (1989). The purpose of this was to provide a critical test of (H_2) as an auxiliary assumption of van Rooij and Schulz (2019), however.

In a much-discussed study, Byrne (1989) presented participants with conditional inference problems like, e.g. "If Lisa has an essay to write, then Lisa will study late in the library", along with an additional premise presenting an alternative antecedent, e.g. "If Lisa has some textbooks to read, then Lisa will study late in the library". Applying this idea to our context, we asked participants for probability evaluations in the presence of alternative causes. We did this by first obtaining alternative causes generated by other participants from a pilot study. We then displayed these above the test questions in the present study for participants in the Alternative-Causes condition. The goal was to see whether the Relevance Effect could be replicated even under full knowledge of alternative causes, when the potential cognitive effort of generating alternative causes had been removed.

Experiment 1 thus provides a critical test of the assumptions needed to account for the Relevance Effect based on van Rooij and Schulz' (2019) causal power account of indicative conditionals.

---

account for degrees of acceptability as here and explicitly apply their theory to data from psychological experiments that asked for degrees of acceptability in the form of subjective probabilities. For this reason, we empirically test such an extension of their theory.

Procedures shared by all Experiments

Experiment 1, like all the other experiments reported in this paper, was conducted as an online study testing a large and demographically diverse sample. Participants were sampled over the Internet (via Mechanical Turk) from the USA, UK, Canada, and Australia. Subjects received a monetary compensation for their participation. The following exclusion criteria were used: 1) not having English as native language, completing the task in less than *min* seconds or in more than *max* seconds,[8] 2) failing to answer two simple SAT comprehension questions correctly in a warm-up phase, 3) answering 'not serious at all' to the question 'how serious do you take your participation' at the beginning of the study, and 4) answering "yes" to whether they recognized the animation from the computer game "Incredible Machines".[9] For each experiment, it was found that these exclusion criteria had a minimal effect on the demographic variables.

To reduce the dropout rate during the experiment, participants first went through three pages in all the experiments. These three pages stated our academic affiliations, posed the two SAT comprehension questions in a warm-up phase, and presented a seriousness check asking how careful the participants would be in their responses (Reips, 2002). Participants were also shown two dummy probability questions to familiarize them with the use of a slider.

Participants

A total of 1004 people completed Experiment 1. After applying the *a priori* exclusion criteria the final sample consisted of 681 participants. Mean age was 39.82 years, ranging

[8] Due to differences among the tasks, the min and max varied between experiments: Experiment 1 = [60s, 1800s], Experiments 2, 3 = [240s, 3600s], Experiment 4 = [120s, 1800s], Experiments 5, 6 = [240s, 1800s].

[9] This last exclusion criterion was used only in Experiments 3-6, which introduced a blackbox condition that required controlling the background knowledge of the participants.

from 18 to 79.[10] 46.1 % of the participants were male. 72.39 % indicated that the highest level of education that they had completed was an undergraduate degree or higher.

Design

The experiment had a between-subjects design with three factors. The first was Relevance (with two levels: Positive Relevance (PO) vs. Irrelevance (IR)). The second was Priors (with four levels: HH vs. HL vs. LH vs. LL; for example, HL means that P(A) = high and P(C) = low). The third was Group (with two levels: Alternative-Causes vs. Control). Thus, there were 16 between-subjects conditions in total.

We will abbreviate the 2 Relevance × 4 Prior conditions as follows below: POHH, POHL, POLH, POLL, IRHH, IRHL, IRLH, IRLL. The Relevance and Prior factors were combined factorially to ensure that the examined relationship generalize across a wide range of different probabilities. This ensures that our results do not merely pertain e.g. to conditionals with high antecedent and consequent probabilities, which tend to sound more plausible, but generalize across a wider spectrum.

Materials and Procedures

Each of the 16 between-subjects conditions was randomly assigned to one of 12 scenarios. Random assignment was performed with replacement, such that each participant saw a different scenario for each condition. This ensured that the mapping of condition to scenario was counterbalanced across participants. One of the 16 between-subjects conditions was randomly assigned to a participant within a block. The block consisted of one page displaying a scenario and three pages presenting the dependent variables (see below). As a reminder, the scenario was presented in grey on the top of these three pages. These scenario texts have been found in previous experiments (Skovgaard-Olsen, Singmann, et al., 2016, 2017) to reliably induce assumptions about relevance and prior probabilities of the antecedent

[10] One participant indicated the age of '14', but given Amazon's regulations we doubt this value.

and the consequent that implement our experimental conditions. Table 3 displays sample items of the Paul scenario for Positive Relevance ($\Delta p > 0$), and Irrelevance ($\Delta p = 0$).

Table 3. Stimulus Materials of the Paul Scenario

Scenario	Paul is driving on a straight road with hardly any traffic ahead. He is on his way to work in an investment bank and is running late. At this point the drive will take about one hour and he is supposed to arrive in 40 minutes.	
	Positive Relevance	**Irrelevance**
HH	If Paul pushes down the gas pedal, then the car will speed up.	If Paul is wearing a shirt, then his car will function normally.
HL	If Paul drives fast, then he will be there in time for work.	If Paul is wearing a shirt, then his car will suddenly break down.
LH	If Paul's car suddenly breaks down, then he will be late for work.	If Paul is wearing shorts, then his car will function normally.
LL	If Paul pushes down the brake pedal, then the car will slow down.	If Paul is wearing shorts, then his car will suddenly break down.
	Positive relevance (PO): mean ΔP = .32 Irrelevance (IR) mean ΔP = -.01	High antecedent: mean P(A) = .70 Low antecedent: mean P(A) = .15 High consequent: mean P(C) = .77 Low consequent: mean P(C) = .27

Note. HL: P(A) = High, P(C) = low; LH: P(A) = low, P(C) = high. The bottom rows display the mean values for all 12 scenarios pretested in (Skovgaard-Olsen, Singmann, et al., 2017). $\Delta p = P(C \mid A) - P(C \mid \neg A)$

For the Paul scenario text in Table 3, participants assume that the event "Paul pushes down the gas pedal" raises the probability of the event "the car will speed up". They moreover assume that both sentences have a high prior probability (Positive Relevance, HH). Conversely, participants assume that the event "Paul is wearing a shirt" is irrelevant for whether "his car will function normally", and that both have a high prior (Irrelevance, HH). Previous studies have moreover confirmed that participants view "Paul pushes down the gas pedal" as a *reason for* the event "the car will speed up" and "Paul is wearing a shirt" as neither a reason for nor against "his car will function normally". The full list of scenarios can be found at: https://osf.io/j4swp/.

On the three randomly ordered pages following the initial scenario, participants were asked to provide estimates of conditional probabilities ($P(C|A)$, $P(C|\neg A)$) via the Ramsey Test. They were thus asked to suppose that the antecedent is the case, and evaluate the

probability of the consequent under this assumption on a scale from 0-100%. In addition, participants were asked to assign probabilities on the same scale to conditional statements across relevance conditions, e.g.: "IF Paul pushes down the gas pedal, THEN the car will speed up".

In a pilot study, we had participants generate alternative causes for the Positive Relevance and Irrelevance items. Two independent raters coded how many independent and plausible causes the participants listed (see https://osf.io/fa9rj/ for the coding instructions). It was found that the rank order $|\text{Alternatives}_{PO}| > |\text{Alternatives}_{IR}|$ obtained not only for the averaged ratings across conditions ($\overline{Alternatives}_{PO}$= 3.13, $\overline{Alternatives}_{IR}$= 2.06), $t(3.56) = 3.39$, $p = 0.033$, but also for each condition and each rater within each condition. For participants in the Alternative-Causes Group, an alternative cause generated by the participants in the pilot study was selected for each of the 96 Relevance × Prior × Scenario combinations. This alternative cause was presented to participants as the antecedent of a conditional. For instance, some participants in the Alternative-Causes Group were shown the following conditional presenting an alternative antecedent for the item above:

"IF Paul is driving down a hill, THEN Paul's car will speed up."

This conditional was displayed on a separate page after the scenario and repeated on every page above the test question for participants in the Alternative-Causes Group. In contrast, participants in the Control Group were presented with the three dependent variables without alternative antecedents.

Results and Discussion

Causal power was calculated based on participants' responses to the conditional probability questions through calculations of ΔP and the following formulas:

[Eq5.] $$power = \begin{cases} \frac{\Delta P}{1-P(C|\neg A)} & if\ \Delta P \geq 0 \\ \frac{-\Delta P}{P(C|\neg A)} & if\ \Delta P < 0 \end{cases}$$

The formulas calculate causal power for generative and preventive causes, respectively.[11] The first goal of the analysis was to establish whether the contrast between the Alternative-Causes and the Control Group influenced the Relevance Effect.

A mixed ANOVA was first conducted using the R-packages `afex` (Singmann et al. 2020) and `emmeans` (Lenth, 2020). The Condition factor (POHH vs. POHL vs. POLH vs. POLL vs. IRHH vs. IRHL vs. IRLH vs. IRLL) and Alternatives factor (Alternative-Causes vs. Control Group) were specified as varying between-subjects. The DV factor (P(C|A) vs. P(C|¬A) vs. P(if A, then C) vs. ΔP vs. power) was specified as a within-subject factor. Through this model, we tested the impact of the Alternative-Causes vs. Control Group contrast on both the three measured (P(C|A), P(C|¬A), P(if A, then C)) and the two calculated dependent variables (ΔP, power) across the between-subjects conditions.

Table 4. ANOVA Table for Experiment 1

Effect	*df*	MSE	*F*	η^2_G	*p*
Condition	7, 665	0.19	73.58	.23	< .0001
Alternatives	1, 665	0.19	2.62	.002	ns
Condition:Alternatives	7, 665	0.19	1.55	.006	ns
DV	2.49, 1655.62	0.12	192.73	.15	< .0001
Condition:DV	17.43, 1655.62	0.12	22.24	.13	< .0001
Alternatives:DV	2.49, 1655.62	0.12	0.79	.0007	ns
Condition:DV:Alternatives	17.43, 1655.62	0.12	0.89	.006	ns

Note. η^2_G is generalized eta squared, which is an effect size measure that is recommended for repeated measures ANOVA in Bakeman (2005). The Alternatives factor encodes the contrast between the Alternative-Causes Group and the Control Group (alternative causes absent).

Given that the contrast between the Alternatives-Causes and the Control Group was neither involved in a simple effect nor in any statistically significant interactions (Table 4), Figure 5 displays the results without this factor:

[11] When ΔP = 0 causal power was stipulated to be zero to avoid the problem of undefined values for cases when P(C|¬A) = 1. Removing the 41 participants with undefined values does not change the relative fit of the models, however. For the purpose of predicting P(if A, then C) by causal power (see M1 below), it would also have been possible to only apply the causal power formula to the subset of cases where ΔP ≥ 0. Figure 5 reveals, however, that the fit of M1 would not have improved by predicting P(if A, then C) = 0 in such cases due to zero generative, causal power.

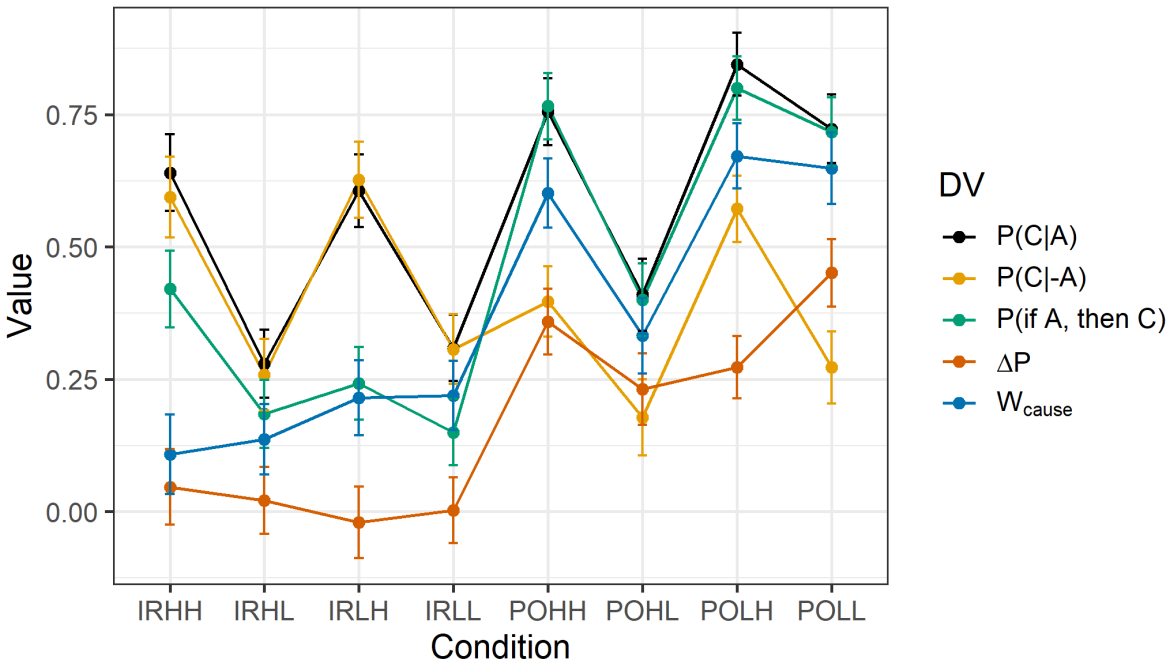


Figure 5. The measured and calculated mean estimates of the five DVs are displayed across the eight Relevance × Priors conditions. The error-bars represent 95% CI intervals.

The systematic differences between P(C|A) and P(if A, then C) for the IR items are noteworthy in Figure 5, because they violate [Eq.1]. At the same time, the two constructs are nearly identical for the PO items. Both findings are in line with the predictions of the Relevance Effect shown in Figure 1. The lack of coincidence of W_{cause} and P(C|A), and the finding that P(C|¬A) estimates are consistently above 0, is also noteworthy, because it casts doubt on van Rooij and Schulz's (2019) auxiliary assumption.

The goal of the second analysis was to test whether causal power predicted P(if A, then C) better than other models. Three mixed linear models were contrasted for modelling P(if A, then C), with random intercepts for scenarios using the R-package `lme4` (Bates et al., 2015):

(M1) a model that predicts P(if A, then C) based on causal power (van Rooij & Schulz, 2019), which measures P(effect|cause,¬alternatives).

(M2) a model that predicts P(if A, then C) by P(C|A) as measured by the Ramsey Test, which corresponds to the suppositional theory of conditionals (Evans & Over, 2004; Oaksford & Chater, 2007; Pfeifer & Kleiter, 2009).

(M3) a model that predicts P(if A, then C) based on an interaction between P(C|A) and the Relevance Condition factor (Positive Relevance vs. Irrelevance), which corresponds to the model used by Skovgaard-Olsen, Singmann, et al. (2016).

The outcome of the model comparison is displayed in Table 5:

Table 5. Model Comparison for Indicative Conditionals

Model		χ^2	df	*p*	AIC	BIC
M1	Causal Power	241.52	1	< .0001	481.70	499.80
M2	P(C|A)	652.26	1	< .0001	232.43	250.52
M3	P(C|A)	515.81	1	< .0001	38.87	66.01
	Relevance Condition	200.67	1	< .0001		
	P(C|A): Relevance Condition	28.72	1	< .0001		

Note. The lower AIC and BIC values indicate that M3 is superior to M1-M2 in light of the parsimoni vs. fit trade-off. 'Relevance' is a categorical factor encoding 'Positive Relevance' vs. 'Irrelevance'.

The information criteria clearly converge on M3. This model permits an interaction between P(C|A) and the Relevance Condition factor such that a lower slope of P(C|A) is expected in the Irrelevance Condition.

In this experiment, the Relevance Effect reported in Skovgaard-Olsen et al. (2016) was replicated both in the Alternative-Causes and the Control Group. It was thereby found that there was no significant effect of explicitly presenting alternative causes to the participants in the manner of Byrne (1989) for the Relevance Effect. This finding, in turn, challenges the auxiliary assumption (H_2) in van Rooij and Schulz (2019) that participants' tendency to ignore alternative causes accounts for the Relevance Effect.

Summary

Based on the pilot study, we know that participants *can* generate alternative causes for both the positive relevance and irrelevance items. Hence, the stimuli in Skovgaard-Olsen, Singmann, et al. (2016) implicitly manipulate the presence of alternative causes. When comparing participants' probability assignments when the presence of alternative causes is

implicitly manipulated (the Control Group) and when it is explicitly manipulated (the Alternative-Causes Group), we find no significant differences (see Table 4).[12]

In a direct model comparison, it was found that when comparing the Suppositional Theory of Conditionals (Evans & Over, 2004; Oaksford & Chater, 2007; Pfeifer & Kleiter, 2009), the causal power theory of the acceptability of indicative conditionals (van Rooij & Schulz, 2019), and the model used in Skovgaard-Olsen, Singmann, et al. (2016), the latter turned out to be the best fitting model. What allowed this model to outperform the other models was that it includes a simple effect of Relevance and an interaction between P(C|A) and the Relevance Condition factor. This interaction term expects a lower slope of P(C|A) in the Irrelevance Condition, where indicative conditionals are predicted to appear defective. At the same time, it allows the use of P(C|A) as a predictor of P(if A, then C), which is especially well-supported in the Positive Relevance Condition. In Appendix B, we further investigate the issue of why causal power theories could not account for our findings through a simulation analysis.

Experiment 2

Beginning with Experiment 2, we used the animated mouse-wheel-machine paradigm.
In Experiment 1, the Relevance Effect was replicated with verbal scenarios. The purpose of Experiment 2 was to replicate this effect using a trial-by-trial learning paradigm involving mechanistic knowledge for the first time.

Method

Participants

[12] As such, the relationship between the Alternative-Causes Group and the Control Group can be viewed as resembling the relationship between the so-called *explicit* paradigm in Byrne (1989) and the *implicit* paradigm in Cummins, et al. (1991). These two paradigms also led to similar results.

A total of 350 people completed the experiment. The same sampling procedures and exclusion criteria were used as in Experiment 1. The final sample after applying the *a priori* exclusion criteria consisted of 221 participants. Mean age was 40.27 years, ranging from 20 to 74. 38.91 % of the participants were male. 69.23 % indicated that their highest level of education was an undergraduate degree.

Design

The experiment had a within-subject design with Relevance as a within-subject factor (with three levels: Positive Relevance (PO) vs. Negative Relevance (NE) vs. Irrelevance (IR)), which refers to three types of items explained below. In total, 20 trials were shown which implemented the following conditions:

Table 6. Experimental Design

	P(C|A)	P(C|¬A)	ΔP
PO	0.83	0.75	0.08
IR	0.80	0.80	0.00
NE	0.75	0.83	-0.08

Note. Contingencies calculated based on the initial trial, where the mc questions were presented, and the subsequent 19 randomly ordered machine trials.

A pilot study[13] had found that although ΔP in the trials shown differed modestly, participants were able to arrive at stronger ΔP differences across conditions when processing the items introduced below. Their background knowledge and the evidence presented concerning the mechanism permitted them to arrive at stronger subjective ΔP values than what was displayed in the trials. These subjective ΔP values correlated with participants' ordinal reason relation assessments, $r_{\text{polyserial}}(97) = .73$, $p < .0001$. The pilot study thus showed that we could use a single contingency condition to reliably manipulate the differences Positive Relevance, Negative Relevance, and Irrelevance using the items introduced below.

Materials and Procedure

[13] https://osf.io/fa9rj/

To ensure that the animations were displayed properly, participants were instructed to adjust their browser so that they would see the whole box in which the animation was presented. We first presented one trial with three multiple-choice questions. After the display of a fixation cross in the upper left corner, participants saw an animation with the mechanistic set-up depicted in Figure 3. In the animation, a blue bowling ball fell down on a mouse-wheel, connected to a conveyor belt, which set a chain of events in action that eventually resulted in a red basketball falling down the basket on the right side of the screen. Participants were instructed that the animations would always start with the display of a white fixation cross in the upper left corner (the position in which the blue bowling ball occurred). Secondly, participants learned that there was a process bar in the middle of the screen that visualizes when the animations would stop. Thirdly, they were asked to pay attention to the animation in all trials, and that they could not press "continue" until all animations had been shown.

In the first trial, the animation paused several times to pose multiple-choice questions to ensure that participants had understood what they had seen. After this trial, participants were given the following instruction:

> As you will see, sometimes the mice can be sleepy ("ZzzZZZz") and fail to run despite being prompted. The mice can also be excited ("Wo hoo!") and start to run without being prompted.

This information was given to make participants aware that (1) the effect could occur in the absence of the target cause and (2) sometimes the effect could remain absent even in the presence of the target cause. The next page informed participants about the change of an irrelevant feature of the machine to implement the Irrelevance condition:

> Sometimes, the bricks also look a bit brighter due to small random shifts in the lights.

Participants then saw 19 further trials implementing the conditions outlined in Table 6. An illustration of the trials can be found on: https://osf.io/fa9rj/.

Following these further animations, three blocks of items were displayed in random order containing several randomly ordered questions. These three blocks implemented the within-subject Relevance factor by presenting participants with the following Positive Relevance (PO), Negative Relevance (NE), and Irrelevance (IR) items, which all concerned properties of the machine shown:

PO: IF the blue bowling ball falls down, THEN the red basketball drops down in the basket.

IR: IF the lights make the bricks in the machine look brighter, THEN the red basketball drops down in the basket.

NE: IF none of the blue bowling balls are moving, THEN the red basketball drops down in the basket.

Within each block, participants were asked to evaluate the probability of these conditionals and the conditional probability of the consequent given the antecedent via the Ramsey Test on a scale from 0% to 100%.

Finally, participants were asked whether they recognized the animation as originating from the computer game "The Incredible Machine", and a list of demographic questions.

Results and Discussion

As a manipulation check, it was found that the following percentages of the participants answered the initial MC question correctly: 81.45%, 96.38%, 94.57%,

Regressing P(if A, then C) on P(C|A), the following differences across Relevance conditions were found:

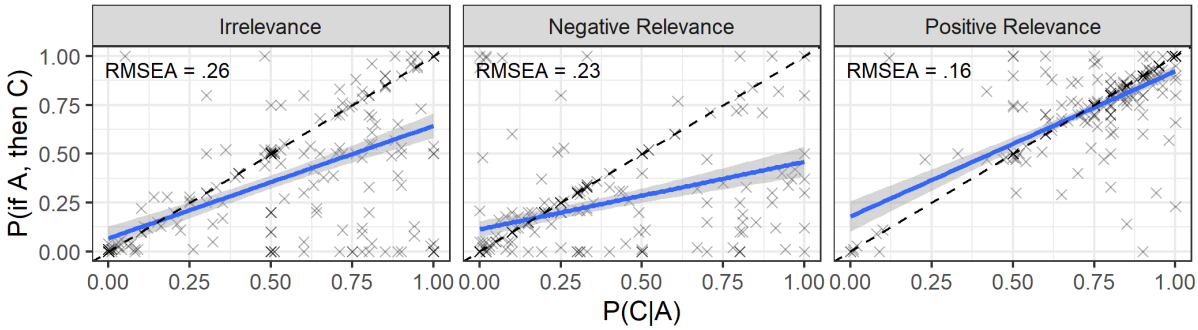


Figure 6. The figure displays predictions of their ratings of P(if A, then C) by their P(C|A) responses. Both variables were rescaled by dividing by 100. The dashed lines indicate the predictions by [Eq. 1]. The root mean square error (RMSEA) values displayed were calculated based on fitting separate least square linear regressions to the PO, NE, and IR conditions.

To test for the influence of Relevance on P(C|A) as a predictor of P(if A, then C), three mixed linear models were contrasted, with random intercepts for participants using the R-package `lme4` (Bates et al., 2015), as shown in Table 7:

Table 7. Model Comparison for Indicative Conditionals

Model		χ^2	df	*p*	AIC	BIC
M1	P(C|A)	637.15	1	< .0001	57.03	75.01
M2	P(C|A)	291.66	1	< .0001	-72.58	-45.60
	Relevance Condition	166.98	2	< .0001		
M3	P(C|A)	302.04	1	< .0001	-86.08	-50.11
	Relevance Condition	172.46	2	< .0001		
	P(C|A): Relevance Condition	24.81	2	< .0001		

Note. Note that 'P(C|A)' here refers to the values measured by the Ramsey Test. The lower AIC and BIC values indicate that M3 is superior to M1 and M2 in light of the parsimoni vs. fit trade-off.

The information criteria favor M3. The results thus indicate that there was both a simple effect of Relevance on P(If A, then C) and an interaction between P(C|A) and Relevance.

For the PO item, the estimated marginal means of the P(if A, then C) ratings were 0.55, 95% CI [0.51, 0.60], 0.74, 95% CI [0.71, 0.77], and 0.92, 95% CI [0.88, 0.97], when P(C|A) was held fixed as 0.50, 0.75, and 1.00, respectively. In contrast, when P(C|A) was held fixed at the same values for the IR item, the estimated marginal means of the P(if A, then C) rating were 0.36, 95% CI [0.33, 0.39], 0.50, 95% CI [0.46, 0.54], and 0.64, 95% CI [0.59, 0.70], respectively. For the NE item, the corresponding values were 0.29, 95% CI [0.25, 0.32], 0.37, 95% CI [0.32, 0.42], and 0.46, 95% CI [0.39, 0.53].

There is a striking match between the data pattern in Figure 6 and the pattern outlined in Figure 1. The results indicate that while participants' responses are well described by [Eq. 1] for the Positive Relevance item, substantial divergences are found for the NE and IR items. Previously, this effect has only been reported using verbal scenarios (Skovgaard-Olsen, Singmann, et al., 2016; Skovgaard-Olsen, Kellen, et al., 2019; Vidal & Baratgin, 2017), which was replicated in Experiment 1. Now we show that this Relevance Effect can also be found in a trial-by-trial learning paradigm in the presence of mechanistic knowledge for the first time.

Experiment 3

To investigate the impact of mechanistic knowledge, Experiment 3 introduced a contrast between one group of participants seeing the underlying mechanism (as in Experiment 2) and another group of participants seeing the same setting covered by a blackbox. The black box concealed the underlying mechanism of the events participants saw (see Figure 4). Our experiments thus allowed us to investigate the effects of knowledge about the operation of a machine, compared with when one can only form associations based on observed covariances in blackbox trials. Experiment 3 used this blackbox manipulation to investigate the impact of participants' causal knowledge on estimates of conditional probabilities and conditional reasoning. Because participants in the blackbox condition only had observed covariances to rely on, we will refer to this group as 'the Regularity Group'.

While other studies have investigated indicative conditionals and singular causation judgments in the same experiment (e.g. Sikorski et al., 2019), we decided to additionally have participants provide counterfactual conditionals and causal power judgments. To investigate the relationship between mechanistic knowledge, causality, conditionals, and contingency, a large online study was therefore conducted with 32 between-subjects conditions that factorially varied these factors.

According to (H_3), causal relations encode multiple conceptual layers, some of which require answers to queries that go beyond what is expressed by indicative conditionals. On the opposing view, indicative conditionals themselves express causal relations. To corroborate (H_3), it would have to be shown that there are aspects of causal relations that go beyond the acceptance of indicative conditionals. Experiments 4-6 were devoted to this aim. In contrast, evidence against (H_3) would have to show that participants evaluate indicative conditionals equivalently to explicit causal notions like singular causation and causal power. Experiment 3 tested this hypothesis.

Experiment 3 therefore investigated whether experimental manipulations known to influence causal reasoning (i.e. contingency conditions and the Machine vs. Blackbox contrast) had a similar impact on four outcome variables of theoretical interest (the probability of indicative conditionals, counterfactual conditionals, singular causation, and causal power). Secondly, Experiment 3 investigated whether participants evaluated these four variables equivalently, or whether differences between them emerged in support of (H_3). To test this, SEM models were fitted to the data across all 32 conditions. A comparison of these models revealed whether it was possible to constraint the four main DVs to be identical. Of interest for these comparisons was whether indicative conditionals were evaluated as explicit causal constructs such as causal power and singular causation in a between-subjects comparison. Thirdly, Experiment 3 was designed to investigate whether the influence of our experimental manipulations on the four main DVs was mediated by participants' estimations of Ramsey Test conditional probabilities. Fourthly, it was investigated whether this mediational relationship in turn was moderated by reason relation assessments.

Method

Participants

A total of 2211 people completed the experiment. The same sampling procedures and exclusion criteria were used as in Experiment 1 with one addition. Experiment 3 additionally

excluded participants who recognized the set-up from the computer game "The Incredible Machine", because such participants will know the mechanism of the machine even in the blackbox condition. The final sample after applying the *a priori* exclusion criteria consisted of 1472 participants. Mean age was 38.94 years, ranging from 18 to 81.[14] 40.42 % of the participants were male. 70.72 % indicated that their highest level of education was an undergraduate degree.

Design

The experiment had a between-subjects design with three factors: DV_{type} (with four levels: indicative conditional vs. singular causation vs. counterfactual conditional vs. causal power), Contingency (with four levels outlined in Table 8 below: a vs. b vs. c vs. d), and Group (with two levels: Machine vs. Regularity, which differed on whether participants saw the underlying mechanism as in Figure 3 or only the blackbox trials as in Figure 4).

Table 8. Experimental design, contingency conditions

	P(C|A)	P(C|¬A)	ΔP	$W_{Antecedent}$
a	0.75	0.50	0.25	0.50
b	0.25	0.00	0.25	0.25
c	0.25	0.25	0.00	0.00
d	0.75	0.75	0.00	0.00

Note. The contingency conditions were introduced through the first initial trial and consecutive 15 randomly ordered blackbox trials. These were subject to the constraint that the last trial displayed was a <bowling ball, basketball> trial. This was done to enable, e.g., participants to make singular causation judgments about whether the bowling ball caused the basketball to fall down the basket. '$W_{Antecedent}$' = the causal power of the antecedent of the conditionals.

Materials and Procedure

Participants were randomly assigned to one of these 32 between-subjects conditions. To investigate the impact of mechanistic knowledge, we first presented one group of participants (those in the Machine condition) with a trial showing the mechanistic set-up from Experiment 2. Participants in the Regularity Group, by contrast, only saw a blackbox trial.

[14] One participant answered '5'. This answer was excluded from the reported age range.

In the first trial, the animation was paused several times to pose multiple-choice questions to ensure that participants had understood what they had seen.

For the 15 trials that followed, all participants saw 15 blackbox trials (Figure 4) conveying the different contingencies listed in Table 8. Participants in the Machine Group were instructed that the blackbox covered most of the animation with the machine that they had seen on the first trial.[15] Following these trials, participants were shown a block with three dependent variables in random order. Two of the dependent variables were shown to all participants. One of these was the following Ramsey Test question:

> Suppose that the blue bowling ball falls down. [highlighted in blue]
>
> Under this assumption, how probable is the following statement on a scale from 0 to 100%:
>
> The red basketball drops down in the basket. [highlighted in blue]

The second question was an ordinal reason relation assessment on a five-point Likert-scale, where the quoted sentences were highlighted in blue:

> Please indicate the extent to which "the blue bowling ball falls down" is a reason for/against "the red basketball falls into the basket":
>
> A strong reason against; a reason against; neutral; a reason for; a strong reason for.

The third dependent variable was a probability judgment on a scale from 0 to 100% with an item determined randomly based on the chosen between-subjects condition from the following list:

> **Singular causation:**
>
> The blue bowling ball caused the red basketball to drop down in the basket.
>
> **Indicative Conditional:**

[15] Since the mechanism was covered for these trials, participants were never exposed to animations of sleepy or excited mice as disablers and alternative antecedents like in Experiment 2. Moreover, since the IR item from Experiment 2 was not used, the colour of the bricks remained constant throughout.

IF the blue bowling ball falls down, THEN the red basketball drops down in the basket.

Counterfactual Conditional:

IF the blue bowling ball had NOT fallen down, THEN the red basketball would NOT have dropped down in the basket.

Causal Power:

Some instances of the red basketball dropping down in the basket are due to hidden alternative causes. Imagine there are 100 runs of the animation in which no alternative causes are present. Suppose that the blue bowling ball falls down in all of these 100 runs. In how many of them would the red basketball drop down in the basket?

The formulation of the causal power question followed a standard formulation found in the literature on causal judgment (see e.g., Cheng & Lu, 2017; Liljeholm & Cheng, 2009).

Finally, participants were asked whether they recognized the animation as originating from the computer game "The Incredible Machine", and a list of demographic questions.

Results and Discussion

Pilot Study for Experiment 3

We first conducted a pilot study. We here summarize some of its results, because they concern the issue of whether participants ignore alternative causes in our experimental paradigm, which was the auxiliary hypothesis used to explain the Relevance Effect in van Rooij and Schulz (2019). Further results concerning the impact of mechanistic knowledge on changes to contingencies are reported on: https://osf.io/fa9rj.

The pilot study presented participants with two open-ended questions, where participants were requested to list up to seven other alternative causes of the basketball dropping into the basket than the blue bowling ball falling down. An acceptable answer to this question might be that one of the mice started to run on their own volition. Secondly,

participants were asked to explain the mechanism in the black box which makes the basketball fall into the basket. To analyze participants' open-ended responses, we had two raters classify the number of alternative causes to the blue bowling ball falling down. As a proxy for the complexity of the explanations, the two raters also classified the number of functional units in participants' explanations of why the red basketball dropped into the basket. Details on the classification can be found on: https://osf.io/fa9rj.

The Machine Group (M = 4.35, SD = 2.25) produced significantly more functional units in their explanations than the Regularity Group (M = 1.84, SD = 1.13), $t(127.18) = 9.17$, $p < .0001$. Moreover, it was found that the Machine Group (M = 1.21, SD = 1.4) produced significantly more alternative causes than the Regularity Group (M = 0.82, SD = 0.99), $t(153.42) = 2.054$, $p = .042$. In the Machine condition, 39.54% produced zero (plausible) alternative causes. In the Regularity condition, 47.5% of the participants produced zero (plausible) alternative causes. However, these proportions did not differ significantly, $\chi^2(1) = 0.77$, $p = 0.38$. In sum, it was found that the explanations of the Machine Group were more complex, as measured by the number of functional units used in their explanations. Moreover, the Machine Group tended to list more alternative causes than the Regularity Group. However, the two groups did not differ in the frequency with which zero physically plausible, alternative causes were listed, which was found to be high (>39%) in both groups.

Main Study

Participants in the Machine Group were asked three MC questions. In the Regularity Group, two MC questions were presented. As a manipulation check, it was found that the following percentages of participants answered the initial mc question correctly: (Machine Group) 83.70%, 98.10%, 96.20%, (Regularity Group) 88.18%, 88.45%.

Structural Equation Model

To analyze all 32 between-subjects conditions, a structural equation model with four groups (one for each of the main dependent variables, y_j) and moderated mediation was fitted

to the data of all 1472 participants (see Figures 7, 8). Structural equation modelling (SEM) is a generalization of regression models used for causal inference in statistics, which is based on modelling the covariance matrix. SEM moreover permits the estimation of direct and indirect effects of explanatory variables as well as imposing conditional independence constraints from a causal model (Kline, 2016; Shipley, 2016). For our purposes, SEM is suited for identifying the sensitivity of our four main outcome variables to the experimental manipulations while holding other factors fixed. Moreover, we use the SEM model for testing the indirect effects of the experimental manipulations through mediating variables.

Due to the theoretical importance of Ramsey Test conditional probabilities, they were considered as a mediator of our manipulations. In line with previous reseach, the indirect paths through the Ramsey test ($P(C|A)_{DV}$) were furthermore moderated by a qualitative reason relation assessment, $Reason_{DV}$. Across the four groups, the two mediators, P(C|A) and Reason, were modeled in the same way. But the model allowed for differential influence of these on the main outcome variable across the four different types.

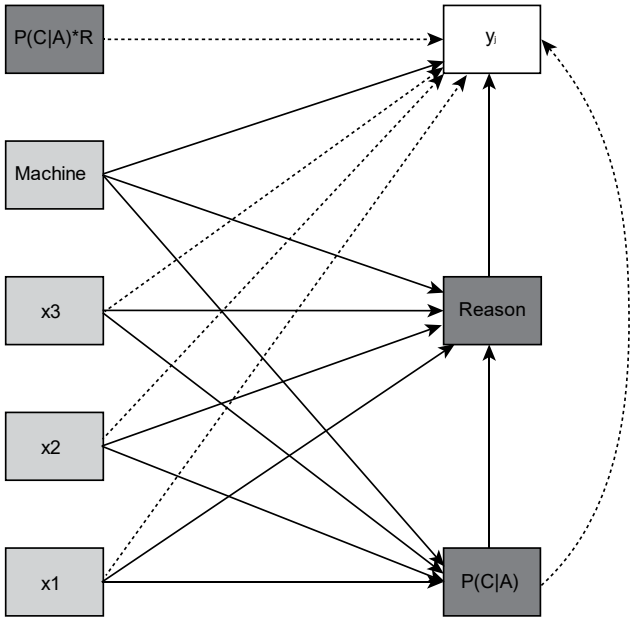


Figure 7. *Conceptual Diagram*. The dashed edges could vary between the four main dependent variables; the solid lines were fixed for all. 'Contingency' (a, b, c, d) was coded into three contrasts: x1, x2, and x3 (see below). A mean structure and covariances (not displayed here) were also added to the SEM model: see https://osf.io/fa9rj for further details. 'P(C|A)*R' = interaction between P(C|A) and Reason.

The model permits the experimental conditions to influence the four main outcomes variables via two causal chains: 1) through the direct effects of the objective input (i.e. the experimental conditions) on the subjectively evaluated DVs, and 2) through indirect effects, where the objective input affects subjective evaluations of P(C|A) and reason relations, which in turn influence the subjectively evaluated DVs. On the hypothesis of a causal interpretation of indicatives, similar psychological processes should be involved in evaluating the four central DVs. The model implements this by allowing the same structure across all four DVs. In addition, the model permits the rejection of this hypothesis by allowing the dashed edges to differ across the four DVs. Comparing models that set the dashed edges equal for some of the four main DVs thus provides a test of differences between these psychological constructs.

In the following, P(C|A) and the four main outcome variables were divided by 100, and the P(C|A) and reason relation were centered on their means. Furthermore, the Contingency factor (a, b, c, d) outlined in Table 8 was encoded in three indicator variables representing the following contrasts: (x1) a - b, (x2) c - b, and (x3) d - b. The model was fitted using the R-package `lavaan` (Rosseel, 2012).

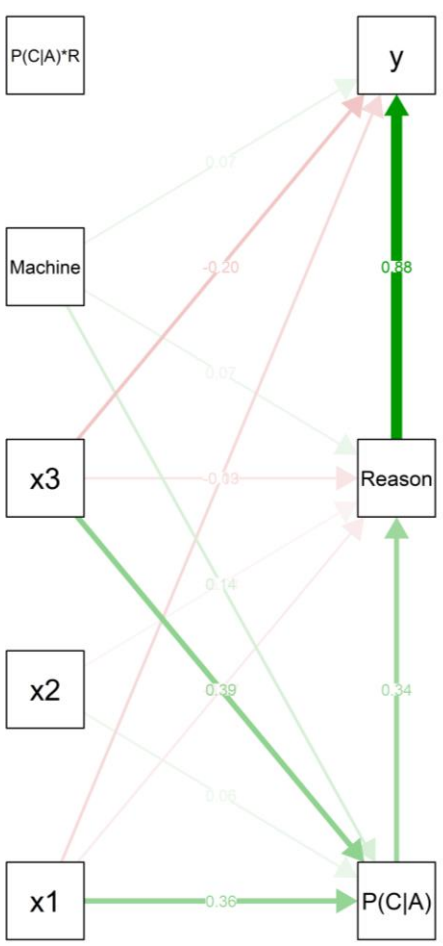
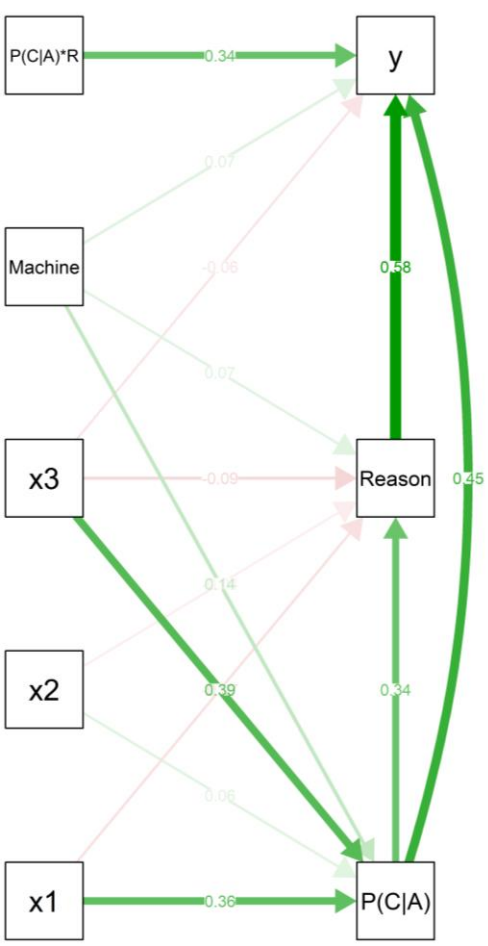
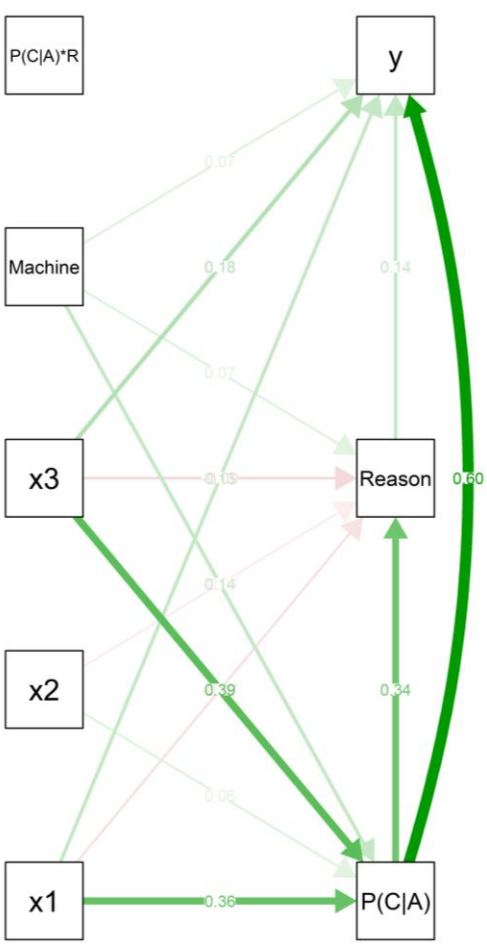
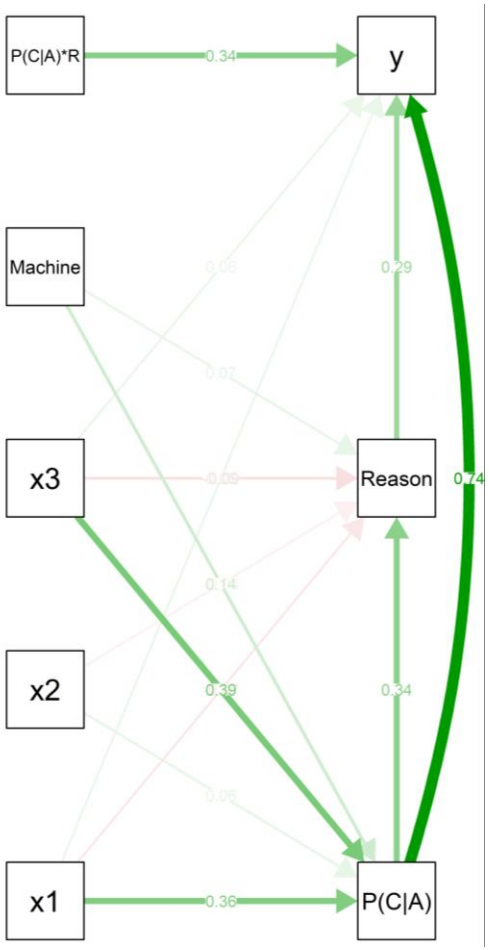
P(counterfactual)	P(singular)	Power	P(indicative)
			
P(C|A) ~ x1 + x2 + x3 + machine, R^2 = 33.3 Reason ~ x1 + x2 + x3 + machine + P(C|A), R^2 = 30.4 *y* ~ x1 + x3 + machine + Reason, R^2 = 31.9	R^2 = 29.9 R^2 = 30.9 *y* ~ x3 + machine + P(C|A)*Reason, R^2 = 54.9	R^2 = 35.7 R^2 = 27.3 *y* ~ x1 + x3 + machine + Reason + P(C|A), R^2 = 63.6	R^2 = 33.7 R^2 = 31.6 *y* ~ x1 + x3 + Reason* P(C|A), R^2 = 82.2

Figure 8. SEM model. 'P(C|A)*R' = two-way interaction of mean-centered P(C|A) and Reason. Contingency contrasts: 'x1' = a – b; 'x2' = c – b; 'x3' = d – b. Only statistically significant effects (p < .05) are shown. The regressions for the two mediators (P(C|A), Reason) are fixed to have the same regression coefficients across groups.

The model in Figure 8 was arrived at by trimming down a saturated model. We did this through a combination of domain knowledge, statistical tests, and by introducing equality constraints between coeffiecients of the predictors of the four main outcome variables. Only statistically significant paths are displayed and were retained. Figure 8 shows that, except for counterfactuals, the linear models of the main outcome variable, y_j, were in each case capable of accounting for more than 50% of the total variance. In the case of indicative conditionals, the model accounted for over 82% of the variance. Global fit statistics moreover indicated that the covariance matrix predicted by the model did not significantly misfit the data, $\chi^2(59) = 73.56$, $p = 0.096$, and that the model met widely used benchmarks for fit measures in SEM modelling (Finch & French, 2015; Kline, 2016): RMSEA = 0.026, 90% CI [0.00, 0.043], $p_{\varepsilon_0 \leq .05} > 0.99$, CFI = .996, SRMR = 0.037, AIC = 1306.85, BIC = 1926.30.

What enabled this model to do comparably well was by imposing differences between the four main DVs corresponding to the dashed edges in the conceptual diagram (Figure 7) and as illustrated in the diagram of the fitted model (Figure 8). In contrast, imposing the constraint that all four main DVs were identical resulted in a model that significantly misfit the data, $\chi^2(68) = 377.18$, $p < 0.001$, and which performed worse in terms of the fit vs. parsimony trade-off, AIC = 1592.47, BIC = 2164.27. Similarly, imposing the constraint that the evaluation of indicative conditionals was identical to causal power and singular causation led to an inferior model that significantly misfit the data, $\chi^2(66) = 175.09$, $p < 0.001$, AIC = 1394.38, BIC = 1976.77. Finally, imposing the constraint that only the evaluation of indicative conditionals and causal power were identical led to an inferior model that significantly misfit the data, $\chi^2(61) = 88.92$, $p = 0.011$, AIC = 1318.21, BIC = 1927.06. Of the latter three, the last was, however, the most competitive. But it still failed to capture the differences between indicative conditionals and causal power displayed in Figure 8.

Across the 16 Contingencies × DV conditions, the main outcome variables were consistently rated higher on the 0-100% scale in the Machine Group than in the Regularity

group ($M_{difference}$=17.57, *SD* = 5.25). Figure 8 shows that this effect was in part mediated through the influence of the Machine factor (0 vs. 1) on the reason relation assessment and the Ramsey test assessment of P(C|A). In addition, Figure 8 shows that all four dependent variables were influenced by the contingency and machine manipulations.

A further finding in Figure 8 is that while the reason relation assessment affected all four main dependent variables to varying degrees, the Ramsey test assessment of P(C|A) did not affect the evaluation of the counterfactual "if A had not been the case, then C would not have been the case". Finally, a moderation of P(C|A) by qualitative reason relation assessments was only found for singular causation judgments and indicative conditionals. We test this moderated mediation effect below.

Ramsey Test Conditional Probabilities and Causal Power

It was found that participants' Ramsey Test conditional probabilities, $P(C|A)_{DV}$, were sensitive to the Machine vs. Regularity manipulation. Participants in the Machine Condition tended to overestimate $P(C|A)_{DV}$ when $P(C|A)_{design}$ = low (conditions: *b*, *c*), $\bar{x}_b$ = .43, $t(185) = 6.67$, $p < .0001$, $\bar{x}_c$ = .48, $t(176) = 9.86$, $p < .0001$. Conversely, participants in the Regularity Condition tended to underestimate $P(C|A)_{DV}$ when $P(C|A)_{design}$ = high (conditions: *a*, *d*), $\bar{x}_a$ = .65, $t(166) = -5.82$, $p < .0001$, $\bar{x}_d$ = .66, $t(201) = -5.55$, $p < .0001$. These divergences from the manipulated conditional probabilities are illustrated through the distances to the dashed lines in Figure 9.

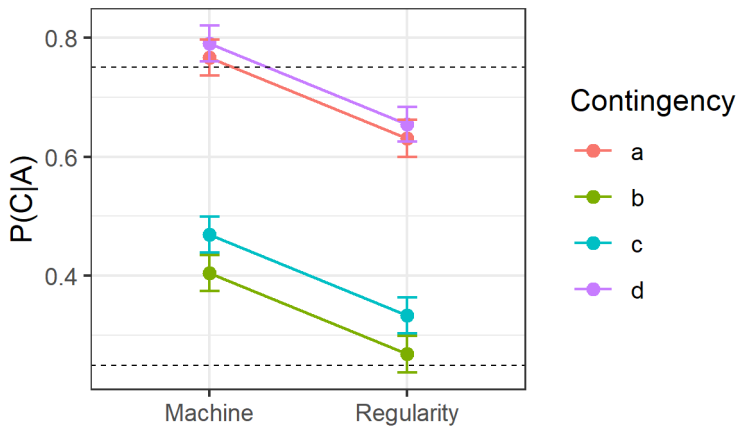


Figure 9. Ramsey Test conditional probabilities across conditions. The dashed lines indicate the manipulated conditional probabilities through the experimental design (see Table 8).

It was, moreover, found that $P(C|A)_{DV}$ was highly correlated with participants' estimates for causal power, $power_{DV}$: $r = 0.90$, $t(367) = 38.56$, $p < .0001$. Controlling for the other predictors shown in Figure 8, $P(C|A)_{DV}$ continued to be a significant predictor of causal power, $b = .60$, $z = 13.65$, $p < .0001$.

The high correlation between Ramsey Test conditional probability and $power_{DV}$ could be interpreted as follows. In the pilot study, it was found that many participants produced zero physically plausible, independent, alternative causes (>39%) in both the Machine and the Regularity conditions when prompted. This finding could in turn be interpreted as supporting van Rooij and Schulz's (2019) hypothesis that participants treat conditional probabilities as equal to causal power because they tend to ignore alternative causes. Such a tendency would count as a bias, insofar as participants also see trials in which the effect occurs in the absence of the target cause. In these trials the effect must be attributed to alternative causes.

However, a comparison with the manipulated conditional probability and causal power through the Contingency conditions invites a different interpretation. Based on $Power_{design}$, the pattern that would be expected for the indicator variables (x1, x2, x3) encoding the Contingency conditions is shown in Table 9:

Table 9. Comparison of causal power and P(C|A)

Indicator	Contingency	$Power_{design}$	Sign	$P(C\|A)_{design}$	Sign
x1	a – b	.50 - .25 = .25	+	.75 - .25 = .50	+
x2	c – b	0 - .25 = -.25	-	.25 - .25 = 0	0
x3	d – b	0 - .25 = -.25	-	.75 - .25 = .50	+

Note. See Table 8 for the Contingency conditions.

As Table 9 shows, the predicted signs of the causal power estimates for the contingency contrasts are: +, -, -. In contrast, the predicted signs for the conditional probability estimates: +, 0, +. Figure 10 displays the signs of participants' causal power

estimates in the data. More specifically, Figure 10 shows the total effects of x1, x2, and x3 on $power_{DV,}$ along with the proportion that is mediated through $P(C|A)_{DV}$ alone:[16]

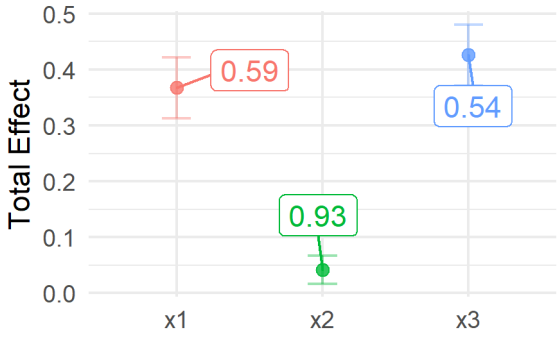


Figure 10. Total effect of x1, x2, and x3 on $power_{DV}$. The proportion of the total (positive) effect that is mediated through $P(C|A)_{DV}$ alone is labelled.

As is clear from a comparison between Figure 10 and Table 9, the magnitudes and signs of the total effects of x1, x2, and x3 on $Power_{DV}$ are more compatible with $P(C|A)_{design}$ than with $Power_{design}$. Had participants estimated $Power_{DV}$ based on the causal power calculated by the actual trials shown, the effects of x1, x2, and x3 would have had to follow the following order: +, -, -. Instead, the effects of x1, x2, and x3 followed the order of the manipulated conditional probabilities: +, 0, +.

The results therefore suggest that the high correlation between Ramsey Test conditional probability and $power_{DV}$ is to be accounted for by participants' causal power estimates as follows: Participants appear to be more sensitive to manipulations in conditional probabilities ($P(C|A)_{design}$) than to variations in the manipulated causal power ($Power_{design}$) in our experimental task. A simulation study in Appendix B shows what the expected relationship is between conditional probabilities and causal power for different types of

[16] Note that there is a slight imprecision in these numbers due to the indirect effects of x1 (*b* = -.009, 95% CI [-.018, -.001]), x2 (*b* = -.006, 95% CI [-.011, .000]), and x3 (*b* = -.013, 95% CI [-.023, -.002]) through the mediator, Reason, with opposite signs. But these adjustments are so slight that they do not impact the interpretation substantially and they are thus ignored in the total (positive) effects displayed below.

statistical analyses. In the General Discussion, we will return to this issue and interpret van Rooij and Schulz's (2019) hypotheses in light of these results.

Moderated Mediation

To test for the influence of reason relation assessments on the indirect effects of Ramsey test conditional probabilities, a moderated mediation analysis was conducted (Hayes, 2018). In this analysis, the Reason factor is used as a moderator of the mediation by the Ramsey Test conditional probabilities. The Reason factor was recoded from its measurement on a five-point Likert-scale to values between 0 and 1: strong reason against (0.2), reason against (0.4), neutral (0.6), reason for (0.8), and a strong reason for (1.0). By trimming down a saturated model, the coefficients were constrainted to be zero for counterfactuals and causal power. Moreoever, as shown in Figure 8, the coeffeicients were set to be equal for singular causation and indicative conditionals. It was found that there was a significant interaction between $P(C|A)_{DV}$ and $Reason_{DV}$ for singular causation and indicative conditionals, $b = .34$, $z = 3.64$, $p < .0001$. In addition, it was found that there was a conditional effect of $P(C|A)_{DV}$ on causal power, $b = .34$, $z = 3.64$, $p < .0001$. Following Hayes (2018), the indirect effect of the experimental conditions through $P(C|A)_{DV}$ on singular causation and indicative conditionals can be viewed as moderated by $Reason_{DV}$, whenever a bootstrap interval of the index of partial moderated mediation does not cross zero (as found in Table 10).

Table 10. Indices of moderated mediation

$X_i \rightarrow P(C\|A) \rightarrow y$		Moderator	Index: a_ib_j	95% Bootstrap CI
$X_i =$	x1	Reason	.12	[.056, .19]
	x2	Reason	.022	[.005, .040]
	x3	Reason	.13	[.060, .21]
	machine	Reason	.047	[.020, .073]

Note. The index 'a_ib_j' is a product out of the regression coefficients of X_i in the mediator regression model (a_i) and the regression coefficients of the moderator on the indirect path (b_j) in the outcome regression model. A bootstrap interval is used, because it has been shown in previous studies that the assumption of normality is violated for this index (Hayes, 2018).

The moderated mediation of P(C|A) by reason relation assessments for the outcome variable, P(if A, then C), replicates the influence of reason relations on P(if A, then C) from Experiments 1 and 2.

Summary

The main findings of Experiment 3 were as follows: First, support for the conceptual layer hypothesis (H_3) could be obtained, because models that treated indicative conditionals and explicit causal constructs (i.e. singular causation and causal power) equivalently were found to significantly misfit the data. Differences between the four main outcome variables thus emerged, which are illustrated in Figure 8. Most notably, it was found that there was no direct effect of Ramsey Test assessments of P(C|A) on counterfactual conditionals ("If A had not been the case, then C would not have been the case") and that the interaction between Ramsey Test conditional probabilities and reason relation assessments could only be found for indicative conditionals and singular causation judgments. In contrast, no such interaction occurred for causal power judgments. Secondly, it was found that Ramsey Test conditional probabilities and measured causal power were highly correlated. It was considered whether this correlation should be interpreted considering the findings of a pilot study showing that many participants failed to produce physically plausible, independent, alternative causes in both the Machine and the Regularity conditions when prompted. Yet, this interpretation was rejected due to the finding that the causal power judgments deviated strongly from the manipulated causal powers. Instead it was found that participants were more sensitive to manipulations of conditional probability in their causal power judgments than to variations in the manipulated causal power in our experimental task.

The general finding of higher ratings in the Machine condition than in the Regularity condition suggests that participants rely on structural information that go beyond mere observed covariances for the four main outcome variables, when mechanistic knowledge is

available. This is in agreement with previous findings (see e.g. Johnson & Ahn, 2017) but is here also found for indicative and counterfactual conditionals. Since our experimental task had this knowledge component, participants had to integrate background information with the observed trials to form their subjective responses (as modelled by the SEM in Figures 7 and 8). But importantly, it was found that participants' evaluations of indicative conditionals could not be equated with their subjective judgments of explicit causal constructs in a between-subjects comparison.

Experiment 4

The results of Experiment 3 displayed in Figure 8 indicate that singular causation and indicative/counterfactual conditionals are influenced by similar factors and mediational processes. Still, it was found that the evaluation of indicative conditionals is not equivalent to the processing of explicit causal notions. According to the hypothesis of multiple conceptual layers of causal understanding (H_3), it is expected that indicative and counterfactual conditionals capture separate components of causal relations. To test this hypothesis, the goal of Experiment 4 was to probe whether participants' singular causation judgments could be predicted by their acceptance of indicative or counterfactual conditionals in a within subject design. This within-subject design was adopted to test whether participants' singular causation judgments could be predicted by their acceptance of indicative and counterfactual conditionals.

Method

Participants

A total of 594 people completed the experiment. The same sampling procedures and exclusion criteria were used as in Experiment 3. The final sample after applying the *a priori* exclusion criteria consisted of 330 participants. Mean age was 40.02 years, ranging from 18 to

74. 46.1 % of the participants were male. 67.88 % indicated that the highest level of education that they had completed was an undergraduate degree or higher.

Procedure

Participants were randomly assigned to the same 4 Contingency × 2 Group between-subjects conditions as in Experiment 3. The procedure was identical to the one in Experiment 3 with one exception: in Experiment 4, only the Singular Causation, Indicative Conditional, and Counterfactual Conditional dependent variables were included, and these were manipulated within subject.

Results and Discussion

To test whether singular causation judgments can be predicted by the probabilities assigned to indicatives and counterfactuals, a within-subject comparison was conducted across the 8 Group (Machine, Regularity) x Contingency (a, b, c, d) conditions. Three regression models were compared (see Table 11 below). First, a model that predicts singular causation judgments based on the Group factor (Machine vs. Regularity) and the probability assigned to indicative conditionals alone (M_1). Secondly, a model that is like (M_1) but additionally includes the probability assigned to counterfactual conditionals as a predictor (M_2). Thirdly, a model that is like (M_2) but additionally controls for the influence of the Contingency factor.

Table 11. Singular Causation Judgments

Model		*b*	SE	*p*	R^2	AIC	BIC
M1	Intercept	.38	.037	< .0001	.30	102.12	117.32
	Indicative	.47	.048	< .0001			
	GroupRegular	-.15	.032	< .0001			
M2	Intercept	.14	.041	< .001	.45	21.61	40.61
	Indicative	.44	.042	< .0001			
	Counterfactual	.38	.040	< .0001			
	GroupRegular	-.076	.029	.0089			
M3	Intercept	.18	.051	< .001	.46	24.67	55.06
	Indicative	.41	.052	< .0001			
	Counterfactual	.39	.043	< .0001			
	GroupRegular	-.077	.029	.008			
	Contingencyb	-.064	.045	ns			
	Contingencyc	-.026	.043	ns			

Contingencyd	-.050	.038	ns

Note. The lower AIC and BIC values indicate that M2 is superior to M1 and M3 in light of the parsimoni vs. fit trade-off.

The model comparison favors (M_2). It was thus found that a model that includes participants' evaluations of counterfactuals was a better fitting model than one that only included indicatives (M_1). This suggests that both the ratings of indicative and counterfactual conditionals were needed to predict singular causation judgments. It was also found that including ratings of counterfactuals accounted for unique variance when including a model that controls for the influence of the experimental conditions (M_3). Thus, even if we take differences in presented contingencies into account, the relationship between singular causation judgments and indicative and counterfactual conditionals holds.[17]

Pearl (2000) and Pearl and Mackenzie (2018) have argued that there are three types of queries that represent different layers of conceptual understanding of causal relations, which can be expressed via conditionals, as we have seen. Here we have not tested interventions. But the results of Experiment 4 indicate that participants' predictive judgments (expressed via indicatives)—and their counterfactual comparisons (expressed via counterfactuals)—are good predictors of their singular causation judgments. This finding is in line with the hypothesis that there are different layers of conceptual understanding of causal relations that can be expressed by natural language conditionals (H_3).

More broadly, the finding that counterfactual judgments influence singular causation judgements is in line with causality theories from philosophy (Goodman, 1947; Lewis, 1973; Collins, Hall, & Paul 2004), computer science (Pearl, 2009), and statistics (Morgan & Winship, 2018; VanderWeele, 2015) emphasizing the close connection between

[17] To control for random effects due to variation across participants in a mixed regression analysis, trial replications would be needed of the DV factor. This would require presenting participants with multiple machines analog to the mouse-wheel machine in Figures 3 and 4. While such an analysis would be desirable, it goes beyond the limits of the present investigation. Aggregating the data and fitting models corresponding to M_1 and M_2 lead to similar results favoring M_2 over M_1 in both Experiment 3 and 4.

counterfactuals and singular causal relations. Finally, in their accounts of singular causation, Pearl (2009, Ch. 10) and Halpern (2019) both build in counterfactual conditions in agreement with our results.

Experiment 5

The results of Experiment 4 suggest that there is more to the acceptance of a causal relation than the endorsement of indicative conditionals. In Experiment 5, this point was further corroborated through the investigation of a common-cause structure with two correlated effects. The use of such common-cause models permitted us to contrast probabilistic dependencies based on spurious correlations with probabilistic dependencies based on direct causal influence.

To further test the hypothesis (H_3) that reasoners grasp multiple conceptual layers of causal relations, Experiment 5 made a direct comparison of the acceptance of indicatives and non-backtracking,[18] interventionalist counterfactuals. Through the common-cause structure, we investigated the contrast between these two types of conditionals in the presence and absence of direct causal relations relating their antecedents and consequents. In our experimental task, the following common-cause version of the mouse wheel machine was implemented. First, a purple bowling ball drops on a mouse wheel, which sets off two sequences of events. One terminating with a yellow basketball following down. Another terminating with a red basketball falling, as shown in Figure 11 below:

[18] In back-tracking counterfactuals, one engages in abductive reasoning and starts reasoning backwards from, e.g., the non-occurrence of an event to the non-occurrence of its typical cause. When modelling interventions in a causal system, this type of reasoning is blocked in Pearl (2009). Pearl achieves this by the stipulation that the intervention sets a variable to a given value while removing the causal influence of variables that would normally have affected it.

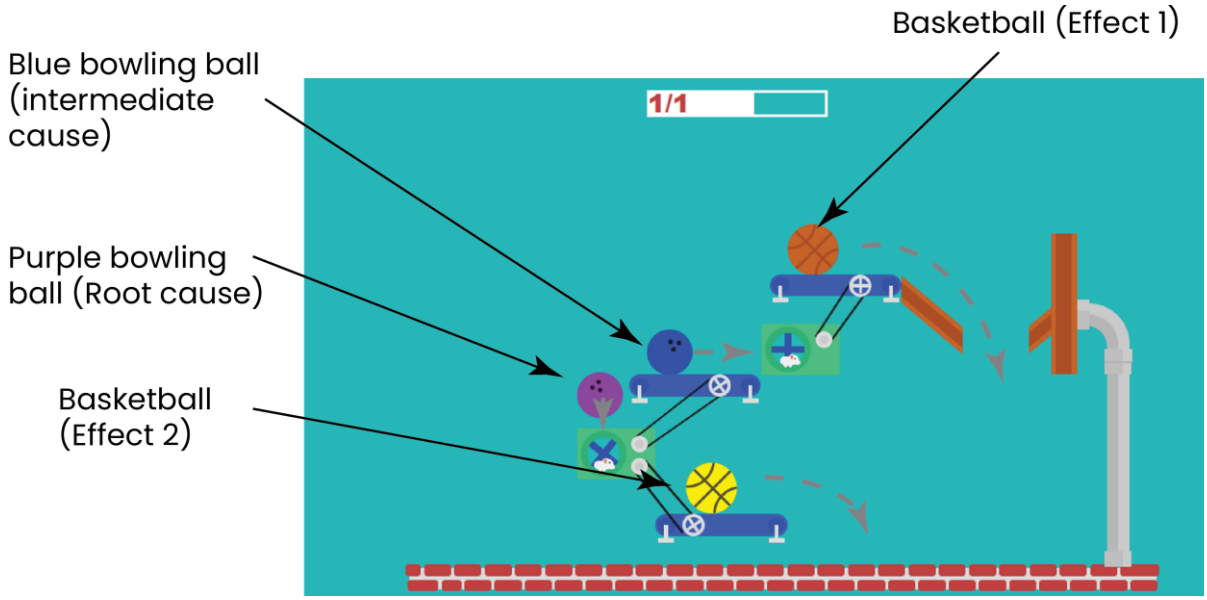


Figure 11. Annotated illustration of a common-cause trial in which the whole mechanism is visible. Instead of the annotation, participants saw animated trials. See https://osf.io/fa9rj/ for a video illustration.

In this common-cause scenario, causal relevance and probabilistic relevance come apart. The reason is that there was a probabilistic dependence between the yellow and the red basketballs, which was not grounded in a direct causal relation. So, although events with the yellow basketball is *relevant* to the probability of the red basketball falling down, the yellow basketball is not a cause for this and is thereby *not causally relevant*.

According to Lassiter (2017), the causal irrelevance of the yellow basketball for the red basketball is decisive for probabilistic counterfactuals. Yet, Lassiter holds that it should play no role for probabilistic indicatives, in line with the following hypotheses:

(H_4) indicatives that *support* and indicatives that *do not support* counterfactuals can be empirically distinguished,

(H_5) the use of indicatives and the acceptance of causal relations can be dissociated even in causal scenarios.

To examine these hypotheses, Experiment 5 contrasts indicatives and counterfactuals in predictive, diagnostic, and common-cause conditions. We moreover compare the assessment of these conditionals with singular causation judgments in situations where the causal relation

is either present or absent. Our goal was to test for possible dissociations between the acceptance of indicative and counterfactual conditionals.

Method

Participants

A total of 949 people completed the experiment. The same sampling procedures and exclusion criteria were used as in Experiment 4. The final sample after applying the *a priori* exclusion criteria consisted of 542 participants. Mean age was 40.16 years, ranging from 18 to 91. 39.48 % of the participants were male. 73.43 % indicated that the highest level of education they had was an undergraduate degree.

Design

The experiment had a mixed design. It contained one within-subject factor, DV (with three levels: indicative conditional vs. singular causation vs. counterfactual conditional). In addition, there were two between-subjects factors: Contingency (with four levels outlined in Table 12 below: a vs. b vs. c vs. d), and Condition (with three levels: predictive vs. diagnostic vs. common-cause). In total, 12 conditions were manipulated between subjects.

Table 12. Experimental design, contingency conditions

	P(E1|C)	$\Delta P_{E1,C}$	P(C|E1)	$\Delta P_{C,E1}$	P(E2|E1)	$\Delta P_{E2,E1}$
a	0.80	0.47	0.80	0.47	0.80	0.47
b	0.50	0.33	0.83	0.33	0.83	0.53
c	0.83	0.33	0.50	0.33	0.80	0.47
d	0.80	0.47	0.80	0.47	0.50	0.33

Note. The contingency conditions were introduced through the first initial trial and consecutive 15 randomly ordered blackbox trials. These trials were subject to the constraint that the last trial displayed was a <bowling ball, yellow basketball, red basketball> trial. This was done to enable participants to make singular causation judgments about whether the bowling ball caused the basketball to fall into the basket.

Materials and Procedure

Participants were randomly assigned to one of the 12 between-subjects conditions. The experimental procedure was similar to the one of Experiment 4. One difference was that Experiment 4 featured a group comparison between the machine vs. blackbox conditions. In

contrast, in Experiment 5 all participants saw an initial common-cause machine trial (see Figure 11) and subsequently 15 blackbox trials, implementing the Contingency conditions outlined in Table 12. Because the common-cause version featured three events, there were eight possible combinations of events. To make the task less complex, participants were instructed in advance which of the three balls they should pay special attention to for answering the questions after the 15 blackbox trials.

Following these trials, participants were shown a block with two types of test questions (the dependent variables) in random order. One of these asked for the probability of an indicative conditional in one of the following three Conditions (predictive vs. diagnostic vs. common-cause), on a slider permitting continuous values between 0-100%.

Predictive Condition:

IF the purple bowling ball falls down, THEN the yellow basketball falls down.

Diagnostic Condition:

IF the yellow basketball falls down, THEN the purple bowling ball fell down.

Two Spuriously Related Effects of a Common-Cause:

IF the yellow basketball falls down, THEN the red basketball drops into the basket.

The second question asked for the probability of a counterfactual conditional. For the counterfactuals, participants were encouraged to imagine an intervention that would have prevented the antecedent from occurring:

> Imagine that we had prevented the purple bowling ball [/yellow basketball] from falling down (e.g. by constructing a safety net under it) [/e.g. by gluing it to the surface].

As a reminder, the statement describing the hypothetical intervention was displayed in grey on the following page. Participants were then asked to rate the probability of one of the following counterfactuals on a scale from 0 to 100% under this assumption:

Predictive Condition:

IF the purple bowling ball had NOT fallen down, THEN the yellow basketball would NOT have fallen down.

Diagnostic Condition:

IF the yellow basketball had NOT fallen down, THEN the purple bowling ball would NOT have fallen down.

Two Spuriously Related Effects of a Common-cause:

IF the yellow basketball had NOT fallen down, THEN the red basketball would NOT have dropped into the basket.

Following this block, participants were asked for singular causation judgments by assigning probabilities to the following statements:

Predictive Condition:

The purple bowling ball falling down caused the yellow basketball to fall down.

Diagnostic Condition:

The yellow basketball falling down caused the purple bowling ball to fall down.

Two Spuriously Effects of a Common-Cause:

The yellow basketball falling down caused the red basketball to drop into the basket.

Results and Discussion

To test whether participants' ratings for the indicative and counterfactual conditional statements were influenced by the Condition and Contingency factors, a mixed ANOVA was fitted to the data. The R-packages `afex` (Singmann et al. 2020) and `emmeans` (Lenth, 2020) were used to this end. Condition (common-cause vs. diagnostic vs. predictive) and Contingency (a vs. b vs. c vs. d) were specified as between-subjects factors. DV (indicative vs. counterfactual vs. singular causation) was specified as a within-subject factor. The goal was to investigate possible dissociations between the probability of indicatives and counterfactuals within the levels of the Condition factor, in line with (H_4) and (H_5).

We found a significant three-way interaction between the Condition, Contingency, and DV factors, $F(11.30, 998.59) = 5.15$, $p < .0001$, $\eta^2_G = .03$. In addition, a significant two-way interaction between the Condition and DV factors was found, $F(3.77, 998.59) = 44.66$, $p < .0001$, $\eta^2_G = .07$. There were also significant simple effects of the DV factor, $F(1.88, 998.59) = 76.34$, $p < .0001$, $\eta^2_G = .06$, and the Condition factor, $F(2, 530) = 123.01$, $p < .0001$, $\eta^2_G = .20$.

The results are displayed in Figure 12 below:

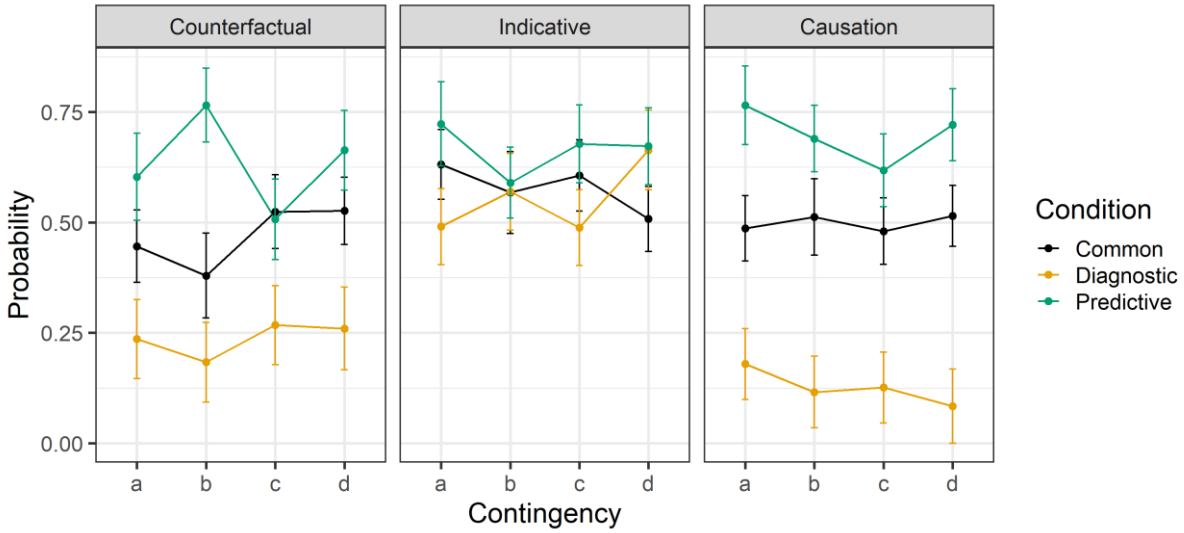


Figure 12. The three DVs are displayed across the 12 levels of the Contingency (a vs. b vs. c vs. d) × Condition (predictive vs. diagnostic vs. common cause) factors. 'causation' = singular causation judgment; 'indicative' = indicative conditional; 'counterfactual' = counterfactual conditional. The error-bars represent 95% CI intervals.

Most participants gave high ratings for the singular causation question in the predictive condition ($M = 0.70$, $SD = 0.24$) and low ratings in the diagnostic condition ($M = 0.13$, $SD = 0.23$). In contrast, they displayed more uncertainty about whether the two target variables were causally linked in the common-cause condition ($M = 0.50$, $SD = 0.32$), with 53 participants in the 3th quantile ($\geq .75$) and 54 participants in the 1th quantile ($\leq .20$).

Bonferroni-Holm corrected pairwise contrasts revealed dissociations between counterfactuals and indicatives. In the common-cause condition, counterfactuals were rated lower than the corresponding indicatives for Contingency a ($b = -0.19$, 95% CI [-0.31, -0.061]), $t(530) = -3.58$, $p < .01$) and Contingency b ($b = -0.19$, 95% CI [-0.33, -0.043]), $t(530) = -3.10$, $p < .01$). For the diagnostic condition, the same relationship was found for

Contingency a (b = -0.26, 95% CI [-0.39, -0.12]), t(530) = -4.52, p < .0001), Contingency b (b = -0.39, 95% CI [-0.52, -0.25]), t(530) = -6.77, p < .0001), Contingency c (b = -0.22, 95% CI [-0.36, -0.086]), t(530) = -3.92, p < .001), and Contingency d (b = -0.40, 95% CI [-0.55, -0.26]), t(530) = -6.84, p < .0001). In the predictive condition, counterfactuals were rated higher than the corresponding indicatives for Contingency b (b = 0.18, 95% CI [0.049, 0.30]), t(530) = 3.32, p < .01) and lower than indicatives for Contingency c (b = -0.17, 95% CI [-0.31, -0.033]), t(530) = -2.97, p < .01).

Across Contingency conditions, it was found, on the one hand, that the three dependent variables were very similar in the predictive condition ($b_{counterfactual - indicative}$ = -0.03, 95% CI [-0.10, 0.038]), t(530) = -1.085, ns; $b_{indicative - causation}$ = -0.033, 95% CI [-0.09, 0.026]), t(530) = -1.34, ns; $b_{counterfactual - causation}$ = -0.064, 95% CI [-0.12, -0.0064]), t(530) = -2.67, p = 0.023). On the other, it was found that the three dependent variables differed increasingly in the common-cause condition ($b_{counterfactual - indicative}$ = -0.11, 95% CI [-0.17, -0.045]), t(530) = -4.09, p < .001; $b_{indicative - causation}$ = 0.080, 95% CI [0.025, 0.13]), t(530) = 3.52, p < .001; $b_{counterfactual - causation}$ = -0.030, 95% CI [-0.08, 0.024]), t(530) = -1.33, ns), and completely in the diagnostic condition ($b_{counterfactual - indicative}$ = -0.32, 95% CI [-0.39, -0.25]), t(530) = -11.06, p < .0001; $b_{indicative - causation}$ = 0.43, 95% CI [0.37, 0.49]), t(530) = 17.64, p < .0001; $b_{counterfactual - causation}$ = 0.11, 95% CI [0.053, 0.17]), t(530) = 4.65, p < .0001).

The results warrant the following conclusions. First, the acceptance of indicatives can clearly become dissociated from the acceptance of the corresponding counterfactuals and singular causation judgments corroborating (H_4) and (H_5). Secondly, it was found that counterfactual judgments tend to align with singular causation judgments. This in turn supports the hypothesis of a hierarchy of causal queries. On this view, singular causation judgments require affirmative answers to counterfactual queries (“does the consequent counterfactually depend on the antecedent?”), in addition to affirmative answers to the

predictive queries ("is the antecedent a good predictor of the consequent?") expressed by indicative conditionals.

The finding of a dissociation was most striking in the comparison between the predictive and diagnostic conditions. A factor contributing to this was the individual variation in whether participants accepted the existence of a direct causal relation in the common-cause condition. The use of blackbox trials may have made it more difficult for the minority who accepted a causal relation in the common-cause condition to distinguish between common-cause conditional and predictive conditionals.

Indicative conditionals can be acceptable both in the direction "if A, then C" and in the direction "if C, then A". This is an indicator that indicatives do not themselves encode causal relations, but rather the inferential potential based on causal (and non-causal) probabilistic dependencies. Whereas causal relations are asymmetrical, our results are consistent with the probabilistic dependency between antecedent and consequents of indicative conditionals being symmetrical (Spohn, 2012a, Ch. 6; Skovgaard-Olsen, 2015).

In Ali et al. (2011), the alternative view is put forward that participants spontaneously recode causal relationships. Accordingly, the consequent can serve as the cause of the antecedent although the reverse direction would normally be expected. This recoding strategy is, however, challenged in cases like the one investigated in Experiment 5, where both the antecedent and the consequent are two effects of a common cause. It is worth noting, moreover, that Ali et al.'s (2011) case for the recoding hypothesis relies on indirect evidence from inference patterns, which showed deviations from participants' responses. We therefore regard this distinction between the spontaneous recoding hypothesis and the hypothesis of symmetry between antecedents and consequents of indicative conditionals (as introduced by the symmetry of probabilistic dependence) as a fruitful area for further inquiry.

Experiment 6

The comparison between predictive and diagnostic conditionals in Experiment 5 involved a tacit comparison between a forward and backward temporal order of the antecedents and consequents. Yet, Experiment 5 only investigated common-cause conditionals in the forward direction, where the event mentioned in the antecedent occurred *before* the event mentioned in the consequent. To exclude possible confounds, Experiment 6 sought to contrast forward and backward common-cause conditionals within a single contingency condition. It was expected that a similar dissociation between indicative and counterfactual conditionals would be found in Experiment 6, and that this dissociation would be moderated by the temporal order (the antecedent occurring *before* vs. *after* the consequent).

Method

Experiment 6 followed the same method as Experiment 5 unless otherwise stated.

Participants

A total of 323 participants completed the experiment. The same sampling procedures and exclusion criteria were used as above. The final sample after applying the *a priori* exclusion criteria consisted of 166 participants. Mean age was 42.86 years, ranging from 19 to 74. 38.55 % of the participants were male. 77.71 % indicated that the highest level of education they had was an undergraduate degree.

Design

The experiment had a mixed design. The DV factor (with three levels: indicative conditional vs. singular causation vs. counterfactual conditional) was varied within subject. The Condition factor was varied between subjects (with four levels: predictive vs. diagnostic vs. common cause forward vs. common cause backward).

In contrast to Experiment 5, only Contingency a of Table 12 was used in Experiment 6. This contingency fixes the conditional probabilities and ΔP values of the examined

conditionals to the same values: P(E1|C) = P(C|E1) = P(E2|E1) = P(E1|E2) = 0.80; $\Delta P_{E1,C}$ = $\Delta P_{C,E1}$ = $\Delta P_{E2,E1}$ = $\Delta P_{E1,E2}$ = 0.47. In total, 4 conditions were manipulated between subjects.

Materials and Procedure

In Experiment 6, the backward common-cause conditional was introduced:

Two Spuriously Correlated Effects of a Common-Cause, Backward:

IF the red basketball dropped into the basket, THEN the yellow basketball fell down.

In addition, Experiment 6 held the tense of all conditionals constant. Both the antecedent and consequents of all examined indicative conditionals were thus manipulated to be in past tense. To create a context of epistemic uncertainty suitable for indicative conditionals, participants were instructed for indicative conditionals that they had to evaluate these sentences with respect to a further unknown run of the animation. For counterfactuals and singular causation judgments, participants were instructed to evaluate the sentences while thinking back on the last trial that they had seen. Like in Experiment 5, this last trial was fixed to be a <bowling ball, yellow basketball, red basketball> trial.

Results and Discussion

An ANOVA with Condition (common-cause backward vs. common-cause forward vs. diagnostic vs. predictive) as a between-subjects factor and DV (indicative vs. counterfactual vs. singular causation) as a within-subject factor was fitted to the data. The R-packages `afex` (Singmann et al. 2020) and `emmeans` (Lenth, 2020) were used. Like in Experiment 5, the goal was to investigate possible dissociations between the probability of indicatives and counterfactuals within the levels of the Condition factor, as a test of (H_4) and (H_5).

It was found that there was a significant two-way interaction between the Condition and DV factors, $F(5.79, 312.79) = 3.33$, $p < .01$, $\eta^2_G = .03$. In addition, significant simple effects of the Condition factor, $F(3, 162) = 23.58$, $p < .0001$, $\eta^2_G = .19$, and the DV factor,

$F(1.93, 312.79) = 9.33$, $p = .0001$, $\eta^2_G = .03$, were found. The results are displayed in Figure 13 below:

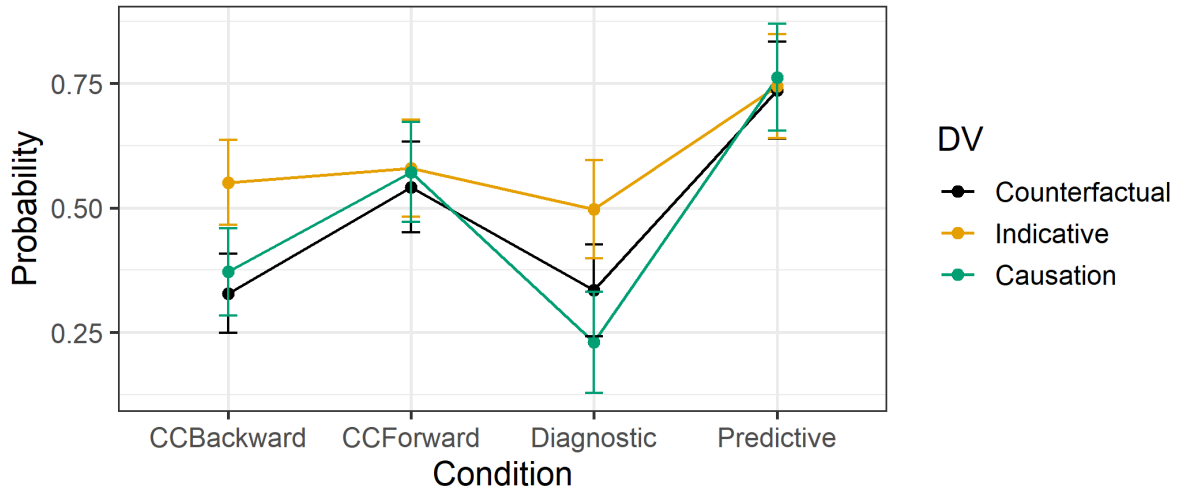


Figure 13. The three DVs are displayed across the 4 levels of the Condition factor. 'CCBackward' = common-cause backward; 'CCForward' = common-cause forward; 'causation' = singular causation judgment; 'indicative' = indicative conditional; 'counterfactual' = counterfactual conditional. The error-bars represent 95% CI intervals.

Bonferroni-Holm corrected pairwise contrasts revealed dissociations between counterfactuals and indicatives. Counterfactuals were rated lower than the corresponding indicatives in the common-cause backward condition (b = -0.22, 95% CI [-0.35, -0.098]), $t(162) = -4.30$, $p = .0001$) and in the diagnostic condition (b = -0.16, 95% CI [-0.31, -0.018]), $t(162) = -2.71$, $p = .015$).

Like in Experiment 5, the results warrant the following conclusions. First, the acceptance of indicative conditionals can become dissociated from the acceptance of the corresponding counterfactuals and singular causation judgments, in accordance with (H_4) and (H_5). Secondly, counterfactual judgments tend to align with singular causation judgments, in line with the hypothesis of a hierarchy of causal queries (H_3). But in contrast to Experiment 5, the dissociation of indicatives and counterfactuals was not found for forward common-cause conditionals. We attribute this difference of results to the procedural changes in Experiment 6, whereby past tense was adopted uniformly for the antecedents and consequents of all indicative conditionals.

One thing is striking about the results shown in Figure 13. Although the conditional probabilities and contingencies were identical for every condition, the indicative conditionals in the predictive condition were systematically higher than in all the other conditions. This finding may have resulted from the participants' need to integrate their background knowledge, and knowledge of the mechanism from the first trial, with their learning experiences in the blackbox trials. Participants may thus have used assumptions about the underlying mechanism to provide clues about how stable the observed covariances were.

Alternatively, the finding could indicate that diagnostic and common-cause conditionals have different acceptability conditions than predictive conditionals. Accordingly, indicative conditionals in the diagnostic and common-cause conditions would have acceptability conditions that are systematically below the corresponding conditional probabilities even under positive contingency. Such a finding would be noteworthy, because it is not part of any of the main theories of indicative conditionals in the psychology of reasoning (see e.g. Bennett, 2003; Evans & Over, 2004; Oaksford & Chater, 2007, 2010a; Rescher, 2007; Douven, 2015; Nickerson, 2015; Goodwin & Johnson-Laird, 2018).

Accordingly, Evans et al. (2007) state that "the Ramsey test predicts that belief in the conditional will be based on the probability of (q|p), regardless of the causal roles instantiated by p and q" (p. 639). To back this up, Evans et al. report evidence concerning predictive and diagnostic conditionals. Worth noticing in their results, however, is that the beta weight does change from .69 in the predictive conditional to .52 in the diagnostic conditional, when the probability of these conditionals is regressed on the corresponding Ramsey test conditional probabilities (see Evans et al. 2007, Table 2). This in turn would be consistent with the hypothesis of different acceptability conditions and the results reported here. Future research will have to determine whether the hypothesis of different acceptability conditions for various types of indicatives is correct in our trial-by-trial learning paradigm and in their paradigm.

General Discussion

The linguistic encoding of knowledge about causal relations plays a vital role for determining the basis for the cultural transfer of causal knowledge across generations. Causative verbs indicating the central contributing factor play a role in this transfer. An example is the verb "to break" in the example "the hammer broke the window" (Neeleman & van de Koot, 2012). Central among the linguistic constructions that facilitate the acquisition of causal knowledge are, moreover, natural language conditionals (Sloman, 2005, Ch. 11; Spohn, 2013). Conditionals play this role as a primary vehicle for expressing dependencies between variables (e.g. 'if you hit it with a hammer, then it will break'). However, exactly which aspects of causal relations are linguistically encoded in indicative conditionals is still very much in dispute; with some authors interpreting recent findings of the role of probabilistic dependency as evidence for a causal interpretation, as we have seen. We will start by discussing what bearing our results have on that debate below and then turn to outlining a more general framework based on Pearl's hierarchical theory of causation in which our various experimental findings can be interpreted in the remainder of the General Discussion.

Indicative Conditional, Causal Power, and the Relevance Effect

Experiment 1 followed previous studies (e.g. Skovgaard-Olsen, Singmann, et al., 2016, Skovgaard-Olsen, Kellen, et al. 2019) in replicating the Relevance Effect with verbal scenarios. In Experiment 2, it was found that the Relevance Effect could also be found in a trial-by-trial learning paradigm involving mechanistic knowledge.

Possible interpretations of the Relevance Effect reported by Skovgaard-Olsen, Singmann, et al. (2016) have played a role in recent work in the psychology of reasoning (see e.g. van Rooij & Schulz, 2019; Oaksford & Chater 2020a, 2020b; Over & Cruz, 2018; Over, 2020). There is a strong temptation to interpret the Relevance Effect as indicating that indicative conditionals are often read causally as that the antecedent is a *cause* of the consequent (Oaksford & Chater, 2020a, 2020b; van Rooij & Schulz, 2019). The latter view

connects with another broad theme; namely, the assumption that causal models underlie most of our subjective judgments of probability (Fernbach et al., 2011). On this view, causal models can thus provide a basic building block for the new paradigm in the psychology of reasoning by, *inter alia*, solving the puzzle of how the Ramsey Test is psychologically implemented. Accordingly, Evans et al. (2007) and Over (2020) both suggest that the Ramsey test is implemented via causal models.

In van Rooij and Schulz (2019), a further step was taken in connecting recent work on the probability of conditionals with theories of causal judgement. van Rooij and Schulz suggest that causal power can be used to account for the acceptability conditions of indicative conditionals (H_1). Since causal power in turn has been used to parameterize causal Bayes nets (Glymour, 2001; Fernbach et al. 2011; Oaksford & Chater, 2017; Aßfalg & Klauer, 2019), this hypothesis would directly show how the subjective probabilities of indicative conditionals could be based on causal models. In addition, van Rooij and Schulz (2019) also suggest as an auxillary hypothesis that participants' tendency to ignore alternative causes could explain why previous research has found evidence in support of [Eq1.] under some conditions (H_2). The reason being that causal power coincides with conditional probability whenever there are no alternative causes.

In line with this conjecture, it was found in the pilot study to Experiment 3 that 39.54% and 47.5% of the participants produced zero (plausible) alternative causes in the Machine condition and the Blackbox conditions, respectively. This finding might in turn explain why causal power and Ramsey test conditional probabilities were found to be highly correlated in Experiment 3 in the trial-by-trial learning paradigm.

To test van Rooij and Schulz's (2019) conjecture (H_2) directly, Experiment 1 made a between-subjects comparison of participants' judgments employing the verbal scenarios originally used to discover the Relevance Effect. In a pilot study preparing such a comparison, it was found, however, that participants had no trouble generating alternative causes for these

stimulus materials both in the Positive Relevance condition and in the Irrelevance condition. In fact, participants tended to generate more alternative causes in the former condition than in the latter. They did this in spite of the fact that the irrelevance items presented participants with a candidate cause (e.g. Paul is wearing a shirt), which was patently useless for producing the effect (e.g. Paul's car suddenly breaking down). To get a more direct critical test, we presented participants with the alternative causes that their peers had generated in a between-subjects comparison in Experiment 1. It made no difference for all the investigated effects whether participants were presented with alternative causes explicitly while making their judgments. These findings suggest that it is not the presence or absence of an accessible alternative cause that accounts for the Relevance Effect.

In a second direct test of van Rooij and Schulz's (2019) conjecture that causal power accounts for the acceptability of indicative conditionals (H_1), it was found in a model comparison in Experiment 1 that neither causal power nor Ramsey Test conditional probabilities alone could account for participants' ratings of P(if A, then C) across conditions. Instead, the analysis replicated Skovgaard-Olsen et al.'s (2016) finding that a model permitting P(C|A) to interact with the Relevance factor best accounted for participants' ratings. In Skovgaard-Olsen, Kellen, et al. (2019) patterns of individual variation in these results were investigated.

Given these negative findings, it is useful to return to the high correlation between causal power and Ramsey test conditional probability in Experiment 3. On closer inspection, it was found that participants' causal power ratings were more sensitive to the manipulated conditional probabilities than the manipulated causal power (see Table 9 and Figure 10). This could suggest that participants were biased in the other direction; by estimating conditional probabilities in a task designed to elicit their causal power judgments. Over et al. (2007, Experiment 2) also found that conditional probabilities calculated based on participants' responses were highly correlated with their ratings of causal strength (r = .87), and that the

latter even correlated with probabilities of conjunctions to the same degree (r = .86). This finding, together with the much weaker associations of causal strength estimates with P(effect|¬cause), could also be interpreted as failures to give proper causal strength estimates in the investigated paradigms.

As a final option, one could adopt a causal power account but drop van Rooij and Schulz's (2019) auxillary assumption that participants' tendency to ignore alternative causes make them evaluate P(if A, then C) as P(C|A). In Appendix B, we investigate this possibility via a simulation analysis. Again, it is found that the simulation analysis did not turn out favorably for a causal power account of P(if A, then C).

Additionally, it was found in Experiment 3 that equating the evaluation of indicative conditionals with judgments of singular causation and causal power would result in a model that significantly misfit the data. In light of these various negative results (as well as further results discussed below), one must be careful not to make the slip from stating that the acceptability of indicative conditionals requires probabilistic dependency to the thesis that indicative conditionals are acceptable just in case there is causal relation between the antecedent and consequent. Instead, our results are consistent with the hypothesis (H_3) that causal relations involve a hierachy of causal queries, which goes beyond what is expressed by indicative conditionals alone.

Having dealt with causal power interpretations of indicative conditionals in relation to debates in the psychology of reasoning, we now turn to our remaining results and broaden our view by outlinning a general framework based on Pearl's hierachical theory of causation in which our various experimental findings can be interpreted.

Learning Causal Relations Through Descriptions

According to Danks (2014): "A full account of causal learning from description remains an open research problem, particularly the question of when learners infer the absence of a causal relation (C does not cause E) from absence of information" (p. 68).

Several of our experiments can be interpreted as providing hints for constructing such an account. In Experiment 4 it was found that singular causation judgments could not be predicted by the probability of indicative conditionals alone. Instead it was found that the probability assigned to both indicatives and counterfactuals was needed to predict singular causation judgments. This finding already suggests that causal relations have multiple conceptual dimensions which are differentially encoded in indicatives and counterfactuals.

In our task involving counterfactuals, participants were asked to evaluate the probability of that the red basketball *would not* have fallen if the blue bowling ball *had not* fallen into the basket, after being shown a trial where both balls fell down. This type of task requires participants to evaluate the following counterfactual probability: $P(Y_{x'} = \text{false} \mid X = \text{true}, Y = \text{true})$. In words: under the assumption that both events actually occurred, what is the probability that Y would not have occurred had X not occurred. According to Pearl (2009), evaluating counterfactual expressions of this type is not possible based on causal Bayes nets, as illustrated in Appendix A. Instead the evaluation of counterfactual expressions requires a causal model with equations that represent in autonomous mechanisms of the data generating processes underlying directed edges, like in structural equation models (SEM), as we explain in Appendix A.

The counterfactual probability evaluates the causal *necessity* of the first event for the second event (*counterfactual query*). In contrast, predictive queries evaluate whether the occurrence of the antecedent is *sufficient for predicting* the consequent. By showing that singular causation judgments cannot be predicted by the acceptance of indicative conditionals alone, our results indicate that participants are sensitive to the counterfactual dimension of causal judgments.

For example, when a colleague says “Germany got the first wave of Covid-19 under control because of masks and social distancing”, and intends a causal interpretation, then this involves accepting the counterfactual, “If Germany had not introduced masks and social

distancing, the first wave of Covid-19 would not have gotten under control". In Spohn (2013, p. 1100), these sentences are taken as equivalent. Our results suggest that the colleague would also have to accept indicative conditionals like "if masks and social distancing are introduced, then Covid-19 will get under control" to make this type of causal attribution. The debate with the colleague over the causal attribution can then be focused on arguments concerning the acceptance/rejection of these indicative and counterfactual conditionals.

Corroborating the hypothesis of differential encoding of multiple conceptual dimensions, it was found in Experiments 5 and 6 that the probability of indicatives and counterfactuals could become dissociated in causal scenarios (H_4). This result was obtained by investigating diagnostic and common-cause conditionals in addition to predictive conditionals. Usually,[19] the focus in the psychology of reasoning has been on the acceptance of predictive, indicative conditionals. Theories have thus been formulated for the probability of indicative conditionals, which do not consider possible asymmetries between the probability of predictive, diagnostic, and common-cause indicative conditionals. Yet such asymmetries were found when holding P(consequent|antecedent) constant in Experiment 6.

Taken together, our finding of the need to predict singular causation judgments based on both indicatives and counterfactuals (Experiment 4) and the dissociations between the latter (Experiments 5 and 6) point in the same direction. They both suggest that one part of an account of causal learning from description may consist in subtle patterns of acceptance and rejection of indicatives and counterfactuals. For instance, the speaker's unwillingness to assert "bad weather would not be coming, if the barometer had been prevented from falling" after having stated "if the barometer falls, bad weather is coming" would suggest that the speaker does not take his/her answer to a predictive query as supporting a causal relation.

[19] One notable exception is Ali et al. (2010, 2011), which complement our results.

Accordingly, the acceptance of an indicative conditional suggests that there is a symmetric, evidential relevance relation between two variables or propositions. But this does not yet imply that the evidential relationship is based on direct causation. As Edgington (2008, p. 18) observes, it is never contradictory to assert 'If A happens, B will happen, but A won't cause B to happen'. In contrast, the acceptance of interventionalist, non-backtracking counterfactuals suggests that there is an asymmetric, direct causal relation. This means that learners should be able to infer the absence of a causal relation from a verbal description indicating either that there is no probabilistic dependency (because the indicative rejected) or that it is a *mere* probabilistic dependency (because the counterfactual is rejected).

Oaksford and Chater (2010b, 2020) have suggested that conditionals describing inferential dependencies can be viewed as structure building operators in causal Bayes nets. The account we have unfolded above is in accordance with this general idea. But the hypothesis of differential linguistic encoding of causal relations through conditionals advanced in this paper opens up for more detailed investigations of the construction of causal models based on linguistic testimony. To illustrate, blackbox observations of three events may either correspond to a causal chain, a common-cause structure, or causal structures with hidden variables. Through indicative conditionals, the edges of the graph can be conveyed. Through the tense of the antecedents and consequents, temporal cues about the ordering of events can be given (e.g. "If it rains, then the streets will be wet" vs. "If the streets are wet, then it rained"). Such temporal cues can be used to infer the direction of edges. Moreover, the acceptance and rejection patterns of interventionalist, non-backtracking counterfactuals can be used to read off the direction of edges. For instance, in a situation where it rains and the streets are wet, "If we had built a pavilion, then the street would not have been wet" sounds acceptable, but "If we had built a pavilion, then it would not have rained" sounds off.

A further component of the ability to infer a qualitative causal structure is the ability to imagine a mechanism whereby cause and effects are related (Lagnado et al., 2007; Johnson &

Ahn, 2017). Our use of the contrast between a Blackbox and a Machine condition led to the finding in Experiment 3 of higher ratings of the four examined outcome variables when mechanistic knowledge was available. This finding suggests that participants rely on structural information that go beyond mere observed covariances when evaluating both conditionals and explicit causal constructs like singular causation and causal power. Assumptions about the underlying mechanism provides clues about how stable observed covariances are and permit participants to make distinctions between predictive/diagnostic relationships and effects of a common cause as in Experiments 5 and 6.

Causal vs. Evidential or Informational Relevance

In Spohn (2010, 2012a, Ch. 14), the distinction between evidential and causal relevance is expressed through the attempt of explicating causal relations as a specific case of a generic reason relation. Pearl (2009) draws a parallel distinction as follows:

> *Informational relevance* is concerned with questions of the form: "Given that we know Z, would gaining information about X give us new information about Y?" *Causal relevance* is concerned with questions of the form: "Given that Z is fixed, would changing X alter Y?" (pp. 234-235, italics added)

The distinction between the evidential and causal relevance of factors also plays a role in distinguishing between purely predictive uses of regression approaches from causally interpreted models in statistics (Gelman & Hill, 2007; Kline, 2016; Pearl, Glymour, & Jewell, 2016; Shipley, 2016; Morgan & Winship, 2018). The distinction is moreover central in discussions over the opposition between evidential and causal decision theory (Hitchcock, 1993; 1996; Meek & Glymour, 1994; Pearl, 2009; Spohn, 2012b).

According to Danks (2014), the graphical models in Bayes nets "can be understood as compact representations of relevance relations, where different types of graphical models present different types of relvance (e.g. informational, causal, probabilistic, communicative)" (p. 39). In causal Bayes nets parameterized via base rates and causal power (see Figure 2), the

directed edges represent relations of *causal relevance*. In contrast, in undirected graphical models, the edges represent symmetric, *evidential relevance* relations (ibid, Ch. 3). In cases of confounding arising via common-cause scenarios, and other cases of spurious correlations,[20] causal relevance and probabilistic relevance come apart. For a psychological theory of probabilistic reasoning, ΔP is often used to represent evidential relevance[21] and causal power can be used to represent causal relevance.

For a causal Bayes net like Figure 2, the parents of a variable represent all the variables that are directly causally relavant to the given variable (Spohn, 2010). Bayes nets are normally only used to encode variables that are at least unconditionally relevant to one another. Answering predictive queries in a Bayes net via conditionalization is therefore unlike the cases of missing-link conditionals, where conditionalization is applied to variables that are categorized as being completly unrelated. Hence, answering predictive queries based on Bayes nets is akin to making predictions based on reason relations.

To acknowledge the counterfactual dimension of causal relations, causal power can also be replaced with the following counterfactual notion of sufficiency: $P(Y_X = \text{true}|Y = \text{false}, X = \text{false})$. In words: under the assumption that both events did not occur, what is the probability that Y *would* have occurred *had X occurred*. This counterfactual concept of sufficiency is identifiable based on Cheng's (1997) account of causal power, provided that no confounding is present and that the cause is generative (Pearl, 2009, ch. 7). The counterfactual notion is, however, stronger than the evidential relationship that we take indicative conditionals to express. The reason is that evidential relevance does not require that the

[20] In addition to spurious correlations created by a common cause, spurious correlations are introduced by conditioning on either a collider or the descendant of a collider in common-effects structures (Pearl et al., 2016).

[21] It should be noted, though, that the factorization of undirected graphical models permits the use of any non-negative function defined over the variables in a clique (Højsgaard et al., 2012), yet ΔP can take negative values. However, ΔP is only one of a larger class of confirmation measures (Crupi et al., 2007) and measures of covariation (Hattori & Oaksford, 2007), which all merit further empirical investigation.

antecedent and the consequent are actually false, but only that the antecedent can be used to *predict* the occurrence of the consequent (as a sufficient reason for believing in the consequent).

Coming from linguistics, Lassiter (2017) puts forward the view that the causal irrelevance of a factor is decisive for probabilistic counterfactuals. At the same time, Lassiter argues that such causal irrelevance plays no role for probabilistic, indicative conditionals. Lassiter argues this point by considering the reversal of truth values of the counterfactual "If Fran had made her flight, it is likely that she would have died". Although this counterfactual would normally be considered true after a plane crash, Lassiter argues that its truth value reverses, when considering the manipulation of the causally relevant factor that Fran is a highly skilled pilot. In contrast, when evaluating the indicative conditional, "If Fran made the flight it is likely that she died", the fact that the plane crashed is held fixed. Varying information about Fran's skills as a pilot should therefore make no difference.[22]

Lassiter's (2017) formal linguistic analyses are in line with Pearl's (2009) idea of a hierarchy of causal queries. They are also congenial to the possibility of mapping natural language expressions of indicatives onto the processing of generic predictive queries and counterfactuals onto the processing of distinctively causal, counterfactual queries, as we have done in the present study. The dissociations of the probability of indicatives and counterfactuals in Experiments 5 and 6 in situations where ratings of singular causation are low corroborate this hypothesis. These dissociations corroborate a conceptual distinction between indicatives that *support* counterfactuals and indicatives that *do not support* counterfactuals (H_4) due to the absence of direct causal relations.

Viewed from this perspective, it is worth highlighting that Kirk (2013) notes in his book on experimental design that scientific hypotheses share the characteristic that they "can

[22] For a dissenting perspective see Over & Cruz (2019) and Over (2020), who hold that counterfactuals can "collapse" to indicative conditionals in examples of this kind.

be reduced to the form of an *if-then* statement. For example, "*If* John smokes, *then* he will show signs of high blood pressure" (p. 49). Kirk proceeds to explain how such if-then statements are to be evaluated through statistical hypothesis testing and confidence interval estimation. But it would have been highly controversial, if he had then gone on to state that these methods of classical statistics were themselves sufficient for establishing causal claims. For this, statistical methods for causal inference make use of procedures for evaluating counterfactuals (Morgan & Winship, 2018; VanderWeele, 2015; Pearl, Glymour, & Jewell, 2016). In addition, the experimental method investigates the scope for intervention, which can now also be emulated through Pearl's (2009) do-calculus based on observational studies.

In other words, validating a scientific hypothesis expressed as an indicative conditional is only the first step towards establishing a causal relation. In addition, it must also be established whether the probabilistic dependency that the conditional expresses can form the basis for intervention, whenever feasible. Secondly, it must be established whether it supports counterfactual conditionals, which are used for causal explanation (e.g. a colleague claiming that "Germany has gotten the first wave of Covid-19 under control *because of* masks and social-distancing"). In short, the assessment of causal relations requires probabilistic *prediction*, investigation of *intervention*, and counterfactually based *explanations*.

The low singular causation ratings in Experiment 6 to the backward common-cause and the backward diagnostic cases further suggest that participants recognize temporal precedence as a requirement of (direct) causal relevance. In both cases, where the antecedent occurred *later* than the consequent, very low singular causation judgments were obtained. Yet, the dissociation of these singular causation judgments with the probability of indicative conditionals suggests that participants accept that the antecedent may nevertheless be evidentially relevant for the consequent, in spite of its low (direct) causal relevance.

Finally, we contrast mental model theory with the account above and make some further comparisons.

Alternative Frameworks

On the newest version of mental model theory (Khemlani et al., 2018), indicatives are viewed as conjunctive assertions about possibilities as shown in table 13:

Table 13. Mapping between indicative and counterfactuals, MMT

Row	Partition		Factual: *If A then C*	Counterfactual: *If A had happened, then C would have happened*	Counterfactual with Neg.: *If A had not occurred, then C would not have occurred*
1	A	C	Possibility	Counterfactual possibility	Fact
2	A	Not-C	Impossibility	Impossibility	Counterfactual possibility
3	Not-A	C	Possibility	Counterfactual possibility	Impossibility
4	Not-A	Not-C	Possibility	Fact	Counterfactual possibility

Note. Quelhas et al. (2018) call indicative conditionals "factual conditionals". The last "Counterfactual with Negations" column was added here.

On this view, 'if the sun is setting, then the sky is red' makes a categorical assertion that it is *impossible* that the sun is setting and the sky is not red, and that it is *possible* that:

the sun is setting and the sky is red,

the sun is not setting and the sky red,

the sun is not setting and the sky is not red.

In Johnson-Laird and Khemlani (2017), various causal relations are also explicated in terms of mental model theory. Interestingly, Johnson-Laird and Khemlani distinguish between a weak and a strong notion of causation. On the weak notion, 'the sun is setting causes the sky to be red' asserts the same three possibilities as 'if the sun is setting, then the sky is red'. The only difference is that the weak notion of causation imposes the temporal constraint that 'the sun is setting' occurs *before* 'the sky is red', whereas indicative conditionals would be compatible with either temporal direction. Hence, on mental model theory, the weak notion of causation is almost identical in meaining to indicative conditionals, but indicative conditionals need not express a causal relation.

In our account, we have emphasized that in addition to accepting indicative conditionals, and respecting a temporal order, counterfactual conditionals of the type 'if the sun had not set, then the sky would not have been red' should be accepted in causal attributions as well. The results from Experiments 4-6 have corroborated this view.

Inspecting Table 13, a special problem emerges for mental model theory in taking this finding on board. The problem is that while the indicative conditional asserts that it is *impossible* that the sun is setting and the sky is not red, the counterfactual with negated antecedent and consequent asserts that this is a *counterfactual possibility*. However, on the notion of impossibility that Johnson-Laird and Khemlani (2017, p. 170) adopt, there exist no possibilities in which an impossible proposition holds. But this means that in accepting an indicative conditional, 'if A, then C', and the counterfactual with negated clauses, 'if A had not occurred, then C would not have occurred', as part of causal attributions, one is depicted as inconsistently claiming *both* that A and not-C is a counterfactual possibility *and* that there are no possibilities in which A and not-C holds. We can therefore conclude that Pearl's hierarchy of causal queries does not sit well with the revised mental model theory.

In philosophy and linguistics, the possible worlds semantics of Stalnaker (1968) and Lewis (1973) remain popular alternatives. Pearl (2009, Ch. 7) showed that it was possible to use his account of interventions in causal models to explicate the elusive notion of similarity in Lewis (1973). In doing so, Pearl showed that it was possible to derive the same conditional logics based on his structural semantics for counterfactuals as on Lewis' account. On this logic, conditional sufficiency, or and-to-if inferences, are valid. For indicative conditionals, these types of inferences are, however, the focus of a recent controversy in the psychology of reasoning (Over & Cruz, 2018; Skovgaard-Olsen, Kellen, et al., 2019).

At the time of Nute (1980), they were already considered problematic for counterfactual conditionals. Accordingly, Nute (1980) discusses various ways of weakening possible worlds semantics into a logic, where they are invalid. Lewis (1973) earlier showed that he could apply his truth conditions for this logic as well if he allowed that other possible worlds could be *as similar* to the actual world as the actual world itself.

In a causal model, this would correspond to considering further possible values of the background variables characterizing the current situation than the ones actually instantiated,

and calculating the effects of forcing the antecedent to be true under those circumstances as well. This could give rise to cases where the consequent is false leading to a failure of conjunctive sufficiency. It would be interesting to see if Pearl could follow the extensions of possible worlds semantics in Nute (1980) to evaluate the counterfactuals with respect to a set of sufficiently similar possible worlds in case the antecedent is true in the actual world.

Other conditional logics have been developed along these lines to avoid conjunctive sufficiency also for indicative conditionals. For instance, Vidal (2017) builds on Nute (1980) but introduces a two-stage impementation of the Ramsey Test that brackets the current beliefs and disbeliefs in the antecedent before evaluating the consequences of adding the antecedent to one's belief set. Simiarly, Rott (2019) has developed a logic for an expanded notion of the Ramsey Test to ensure that the antecedent is relevant for the consequent, which he suggests could either be part of the truth or acceptability conditions of indicative conditionals. See further Raidl (2020) for an overview of several such formal systems.

Conclusion

In sum, the evidence across the six experiments we reported is most consistent with the view that indicative conditionals encode inferential relations (as shown by the Relevance Effect, which was replicated in Experiments 1 and 2) and are used to answer predictive queries. Following Skovgaard-Olsen, Collins, et al. (2019), these inferential relations may be viewed as conventional implicatures. The results also suggest that there are multiple layers of conceptual understanding involved in causal relations that are differentially encoded in indicative and counterfactual conditionals, which has not been demonstrated before. Both the acceptance of indicatives and counterfactuals are required to predict singular causation judgments (Experiment 4). However, when the acceptance of indicative and counterfactual conditionals become dissociated (Experiments 5 and 6), the acceptance of counterfactuals track singular causation judgments and the (direct) causal relevance of the antecedent for the consequent. In contrast, indicative conditionals track evidential relevance.

Moreover, although causal power may be used to parameterize causal Bayes nets (Glymour, 2001), and its application to indicative conditionals can be theoretically motivated (van Rooij & Schulz, 2019), it turns out empirically that causal power does not fit our data for indicative conditionals (Experiments 1, 3, Appendix B). Instead, an account that assumes that participants make reason relation assessments using conditional probabilities while being sensitive to when the antecedent lowers or raises the probability of the consequent turns out to better account for our results. This is in line with the idea of indicative conditionals as answering predictive queries requiring evidential relevance without necessarily representing causal relevance.

References

Ali, N., Schlottmann, A., Shaw, A., Chater, N., and Oaksford, M. (2010). Causal discounting and conditional reasoning in children. In: Oaksford, M. and Chater, N. (Ed.), *Cognition and Conditionals* (pp. 117-134). Oxford: Oxford University Press.

Ali, N., Chater, N., and Oaksford, M. (2011). The mental representation of causal conditional reasoning: Mental models or causal models. *Cognition*, *119*(3), 403-18.

Arlo-Costa, Horacio (2007). The Logic of Conditionals. In E. N. Zalta (eds.), *The Stanford Encyclopedia of Philosophy* (spring 2016 Edition). Retrieved from: <http://plato.stanford.edu/archives/fall2016/entries/logic-conditionals/>.

Adams, E. W. (1975). *The Logic of Conditionals*. Dordrecht: D. Reidel.

Andreas, H., & Günther, M. (2018). A Ramsey Test Analysis of Causation for Causal Models, *The British Journal for the Philosophy of Science*. https://doi.org/10.1093/bjps/axy074

Aßfalg, A., & Klauer, K. C. (2019). Reasoners Consider Alternative Causes in Predictive and Diagnostic Reasoning. *Journal of Experimental Psychology: Learning Memory and Cognition*, 1–61.

Bakeman, R. (2005). Recommended effect size statistics for repeated measures designs. *Behavior Research Methods, 37*(3), 379-384.

Bates, D., Mächler, M., Bolker, B., Walker, S. (2015). Fitting Linear Mixed-Effects Models Using lme4. *Journal of Statistical Software, 67*(1), 1-48.

Bennett, J. (2003). *A Philosophical Guide to Conditionals*. Oxford: Oxford University Press.

Bouton, M. E. (2016). *Learning and Behvaior: A Contemporary Synthesis.* Oxford: Oxford University Press.

Brandom, B. (1994). *Making it Explicit*. Cambridge, Mass.: Harvard University Press.

Byrne, R. M. J. (1989). Suppressing valid inferences with conditionals. *Cognition, 31*, 61-83.

Cheng, P. W., & Lu, H. (2017). Causal Invariance as an Essential Constraint for Creating a Causal Representation of the World: Generalizing the Invariance of Causal Power. In *Waldmann, M. R. (Eds.), The Oxford Handbook of Causal Reasoning* (pp. 65–84). Oxford: Oxford University Press.

Cheng, P. W. (1997). From Covariation to Causation: A Causal Power Theory. *Psychological Review, 104*(2), 367–405.

Collins, J., N. Hall, and L. A. Paul (Eds.) (2004). *Causation and Counterfactuals*. Cambridge, Mass.: MIT Press.

Crupi, V., Tentori, K., and Gonzalez, M. (2007). On Bayesian Measures of Evidential Support: Theoretical and Empirical Issues. *Philosophy of Science, 74*, 229-252.

Cruz, N., Over, D., Oaksford, M., & Baratgin, J. (2016). Centering and the meaning of conditionals. In A. Papafragou, D. Grodner, D. Mirman, & J. C. Trueswell (Eds.), Proceedings of the 38th annual conference of the cognitive science society (pp. 1104–1109). Austin, TX: Cognitive Science Society.

Cummins, D. D., Lubart, T., Alsknis, O., and Rist, R. (1991). Conditional reasoning and causation. *Memory and Cognition, 19*, 274-282.

Cummins, D. (1995). Naive theories and causal deduction. *Memory & Cognition, 23*(5), 646–

658.

Cummins, D. D. (2014). The impact of disablers on predictive inference. *Journal of Experimental Psychology: Learn. Mem. Cogn., 40*, 1638–1655

Danks, D. (2014). *Unifying the Mind. Cognitive Representations as Graphical Models.* Cambridge, Massachusetts: The MIT Press.

Darwiche, A. (2009). *Modelling and Reasoning with Bayesian Networks*. Cambridge: Cambridge University Press.

Douven, I. (2016). *The Epistemology of Indicative Conditionals*. Cambridge: Cambridge University Press.

Edgington, D. (1995). On conditionals. *Mind*, *104*(414), 235–329.

Edgington, D. (2008). I-Counterfactuals. *Proceedings of the Airtotelian Society, 108*(1), 1-21.

Elqayam, S. & Over, D. E. (2013). New paradigm psychology of reasoning: An introduction. In S. Elqayam, J. Bonnefon, and D. E. Over (Eds.), *Thinking & Reasoning, 19* (3-4), 249-265.

Evans, J. St. B. T. (2020). The suppositional conditional is not (just) the probability conditional. In Elqayam, E., Douven, I., Evans, J. St. B. T., and Cruz, N. (Eds.), *Logic and Uncertainty in the Human Mind* (pp. 57-70). London: Routledge.

Evans, J. St. B. T. and Over, D. (2004). *If*. Oxford: Oxford University Press.

Evans, J. St. B. T., Handley, S., Hadjchristidis, C., Thompson, V., Over, D., & Bennett, S. (2007). On the basis of belief in causal and diagnostic conditionals. *The Quarterly Journal of Experimental Psychology*, *60*(5), 635–643.

Evans, J. S. B. T., Handley, S. J., & Over, D. E. (2003). Conditionals and conditional probability. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(2), 321–335.

Evans, J. S. B. T., Over, D. E., & Handley, S. J. (2005). Suppositions, extensionality, and conditionals: A critique of the mental model theory of Johnson-Laird and Byrne (2002). *Psychological Review*, *112*(4), 1040–1052.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2010). Neglect of alternative causes in predictive but not diagnostic reasoning, *Psychological Science, 21,* 329–336.

Fernbach, P. M., Darlow, A., & Sloman, S. A. (2011). Asymmetries in Predictive and Diagnostic Reasoning. *Journal of Experimental Psychology: General*, *140*(2), 168–185.

Fernbach, P.M., and Erb, C.D. (2013).A quantitative theory of conditional reasoning. *Journal of Experimental Psychology: Learn. Mem. Cogn., 39,* 1327–1343

Fernbach, P. M., & Rehder, B. (2013). Cognitive shortcuts in causal inference. *Argument and Computation*, *4*(1), 64–88.

Finch, W. H., French, B. F. (2015). *Latent variable modeling with R*. New York: Routledge.

Gelman, A. and Hill, J. (2007). *Data Analysis using Regression and Multilevel/Hierarchical Models*. Cambridge: Cambridge University Press.

Glymour, C. (2001). *The Mind's Arrows, Bayes Nets and Graphical Causal Models in Psychology*. Cambridge, MA: the MIT Press.

Goodman, N. (1947). The Problem of Counterfactual Conditionals. *The Journal of Philosophy*, *44*(5), 113.

Goodwin, G. P., and Johnson-Laird, P. N. (2018). The Truth of Conditional Assertions. *Cognitive Science, 42*, 2502-2533.

Gopnik, A., Glymour, C., Sobel, D. M., Schulz, L. E., Kushnir, T., & Danks, D. (2004). A Theory of Causal Learning in Children: Causal Maps and Bayes Nets. *Psychological Review, 111*(1), 3–32.

Halpern, J. (2019). *Actual Causality*. Cambridge, MA: the MIT Press.

Hayes, A. F. (2018). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (2. Edition). New York: The Guilford Press.

Hattori, M. and Oaksford, M. (2007). Adaptive Non-Interventional Heuristics for Covariation Detection in Causal Induction: Model Comparison and Rational Analysis. *Cognitive Science, 31,* 765-814.

Hertwig, R., & Erev, I. (2009). The description–experience gap in risky choice. *Trends in Cognitive Sciences, 13*(12), 517–523. https://doi.org/10.1016/j.tics.2009.09.004

Hitchcock, C. (1993). A Generalized Probabilistic Theory of Causal Relevance. *Synthese, 97*, 335-364.

Hitchcock, C. R., (1996). Causal Decision Theory and Decision-Theoretic Causation. *Noûs, 30*(4), 508–526.

Højsgaard, S., Edwards, D., and Lauritzen, S. (2012). Graphical Models with R. New York: Springer.

Johnson, S. G. B., & Ahn, W. (2017). Causal Mechanisms. In M. R. Waldmann (Ed.), *The Oxford Handbook of Causal Reasoning* (pp. 127–146). Oxford: Oxford University Press.

Johnson-Laird, P. N. and Khemlani, S. S. (2017). Mental Models and Causation. In M. R. Waldmann (Ed.), *Oxford library of psychology. The Oxford handbook of causal reasoning* (p. 169–187). Oxford University Press.

Kahneman, D., Slovic, P., & Tversky, A. (1982). *Judgment under uncertainty: heuristics and biases*. Cambridge: Cambridge University Press. In M. R. Waldmann (Ed.), *The Oxford Handbook of Causal Reasoning* (pp. 169–187). Oxford: Oxford University Press.

Khemlani, S. S., Byrne, R. M. J., & Johnson-Laird, P. N. (2018). Facts and Possibilities: A Model-Based Theory of Sentential Reasoning. *Cognitive Science, 42*(6), 1887–1924.

Kern-Isberner, G. (2001). *Conditionals in nonmonotonic reasoning and belief revision : considering conditionals as agents*. Berlin: Springer.

Kirk, R. E. (2013). *Experimental Design. Procedures for the Behavioral Sciences.* London: Sage Publications. (*4th Edition*)

Kline, R. B. (2016). *Principles and practice of structural equation modeling* (4. Edition). New York: The Guildford Press.

Kratzer, A. (2012). *Modals and Conditionals.* Oxford: Oxford University Press.

Krzyżanowska, K., Wenmackers, S., & Douven, I. (2013). Inferential Conditionals and Evidentiality. *Journal of Logic, Language and Information*, *22*(3), 315-334.

Krzyżanowska, K., Collins, P. J., & Hahn, U. (2017). Between a conditional's antecedent and its consequent: Discourse coherence vs. probabilistic relevance. *Cognition*, *164*, 199–205.

Lagnado, D. A., Gerstenberg, T., and Zultan, R. (2014). Causal Responsibility and Counterfactuals. *Cognitive Science*, *37*, 1036-1073.

Lagnado, D. A., Waldmann, W. R., Hagmayer, Y., and Sloman, S. A. (2007). Beyond covariation: Cues to causal structure. In A. Gopnik & L. Schulz (Eds.), *Causal learning: Psychology, philosophy, and computation* (p. 154–172). Oxford: Oxford University Press.

Lassiter, D. (2017). Probabilistic language in indicative and counterfactual conditionals. *Proceedings of SALT, 27*, 525-546.

Lenth, R. (2020). *emmeans: Estimated Marginal Means, aka Least-Squares Means*. R package version 1.5.1. https://CRAN.R-project.org/package=emmeans

Lewis, D. (1973). Causation. *The Journal of Philosophy*, *70*(17), 556.

Lewis, D. (1973). *Counterfactuals*. Cambridge, MA: Harvard University Press.

Liljeholm, M., & Cheng, P. W. (2007). When is a cause the "same"? Coherent generalization across contexts. *Psychological Science, 18*(11), 1014–1021.

Liu, M. (2019). Current issues in conditionals. *Linguistic Vanguard, 5*(3), 1-8.

Luhmann, C. C., & Ahn, W. (2005). The meaning and computation of causal power: A critique of Cheng (1997) and Novick and Cheng (2004). *Psychological Review, 112*, 685–692.

Manktelow, K. (2012). *Thinking and Reasoning: an Introduction to the Psychology of Reason, Judgment and Decision Making.* Hove: Psychology Press.

Mayrhofer. R. & Waldmann, M. (2015). Agents and Causes: Dispositional Intuitions As as a

Guide to Causal Structure. *Cognitive Science, 39,* 65-95.

Meder, B., Mayrhofer, R., and Waldmann, M. R. (2014). Structure Induction in Diagnostic Causal Reasoning. *Psychological Review, 121*(3), 277-301.

Meek, C. and Glymour, C. (1994). Conditioning and Intervening. *The British Journal for the Philosophy of Science, 45,* 1001-1021.

Morgan, S. L. and Winship, C. (2018). *Counterfactuals and Causal Inference* (2th Edition). Cambridge: Cambridge University Press.

Neeleman, A., & van de Koot, J. (2012). The Linguistic Expression of Causation. In M. Everaert, T. Siloni, M. Marelj (Eds.), *The Theta System: Argument Structure at the Interface* (pp. 20-51). Oxford: Oxford University Press.

Nickerson, R. (1998). Confirmation Bias: A Ubiquitous Phenomenon in Many Guises. *Review of General Psychology, 2*(2), 175-220.

Nickerson, R. (2015). *Conditional Reasoning. The Unruly Syntactics, Semantics, Thematics, and Pragmatics of "If"*. Oxford: Oxford University Press.

Nute, D. (1980). *Topics in Conditional Logic*. Dordrecht: Reidel.

Oaksford, M. and Chater, N. (2007). *Bayesian Rationality: The Probabilistic Approach to Human Reasoning*. Oxford: Oxford University Press.

Oaksford, M., & Chater, N. (Eds.). (2010a). *Cognition and conditionals: probability and logic in human thinking*. Oxford: Oxford University Press.

Oaksford, M., & Chater, N. (2010b). Causation and conditionals in the cognitive science of human reasoning. *Open Psychology Journal, 3*, 105-118.

Oaksford, M., & Chater, N. (2017). Causal Models and Conditional Reasoning. In M. R. Waldmann (Eds.), *The Oxford Handbook of Causal Reasoning* (pp. 327–346). Oxford: Oxford University Press.

Oaksford, M., & Chater, N. (2020a). New Paradigms in the Psychology of Reasoning. *Annual Review of Psychology, 71*(1), 1–26.

Oaksford, M., & Chater, N. (2020b). Integrating Causal Bayes Nets and Inferentialism in Conditional Inference. In Elqayam, E., Douven, I., Evans, J. St. B. T., and Cruz, N. (Eds.), *Logic and Uncertainty in the Human Mind* (pp. 116-132). London: Routledge.

Oberauer, K., & Wilhelm, O. (2003). The meaning(s) of conditionals: Conditional probabilities, mental models, and personal utilities. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*(4), 680–693.

Over, D. E. (2020). The Development of the New Paradigm in the Psychology of Reasoning. In Elqayam, E., Douven, I., Evans, J. St. B. T., and Cruz, N. (Eds.), *Logic and Uncertainty in the Human Mind* (pp. 243-263). London: Routledge.

Over, D. E., & Cruz, N. (2018). Probabilistic accounts of conditional reasoning. In L. J. Ball and V. A. Thompson (Eds.), *International handbook of thinking and reasoning* (pp. 434– 450). Hove, UK: Psychology Press.

Over, D. E., & Cruz, N. (2019). Philosophy and the psychology of conditional reasoning. In A. Aberdein & M. Inglis (Eds.), *Advances in experimental philosophy of logic and mathematics* (pp. 225-249). London: Bloomsbury Academic.

Over, D. E., Hadjichristidis, C., Evans, J. S. B. T., Handley, S. J., & Sloman, S. A. (2007). The probability of causal conditionals. *Cognitive Psychology*, *54*(1), 62–97.

Pearl, J. (2009). *Causality: models, reasoning, and inference* (2th Ed.). Cambridge: Cambridge University Press.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems: Networks of Plasuible Inference*. San Francisco: Morgan Kaufman Publishers.

Pearl, J., Glymour, M., and Jewell, N. P. (2016). *Causal Inference in Statistics*. Sussex: Willey.

Pearl, J., & Mackenzie, D. (2018). *The Book of Why: The New Science of Cause and Effect*. New York: Baisc Books.

Pelley, M. E. L., Griffiths, O., and Beesley, T. (2017). Associative Accounts of Causal

Cognition. In Waldmann, M. R. (Eds.), *The Oxford Handbook of Causal Reasoning* (pp. 13–28). Oxford: Oxford University Press.

Pfeifer, N., & Kleiter, G. D. (2009). Framing human inference by coherence based probability logic. *Journal of Applied Logic, 7*(2), 206–217.

Politzer, G., & Bonnefon, J. F. (2006). Two varieties of conditionals and two kinds of defeaters help reveal two fundamental types of reasoning. *Mind and Language*, *21*(4), 484–503.

Quelhas, A. C., Rasga, C., & Johnson-Laird, P. N. (2018). The Relation Between Factual and Counterfactual Conditionals. *Cognitive Science*, *42*(7), 2205–2228.

Raidl, E. Definable Conditionals. *Topoi* (2020). https://doi.org/10.1007/s11245-020-09704-3

Rehder, B. (2014). Independence and dependence in human causal reasoning. *Cognitive Psychology*, *72*, 54–107.

Rehder, B. and Waldmann, M. R. (2017). Failures of explaining away and screening off in described versus experienced causal learning scenarios. *Memory & Cognition*, *45*, 245-260.

Reips, U. (2002). Standards for Internet-Based Experimenting. *Experimental Psychology, 49*(4), 243–256.

Rescher, N. (2007). *Conditionals*. Cambridge, MA.: The MIT Press.

Rosseel, Y. (2012). lavaan: An R Package for Structural Equation Modeling. *Journal of Statistical Software*, 48(2), 1-36.

Rott, H. (1986). Ifs, though, and because. *Erkenntnis*, 25, 345–70.

Rott, H. (2019). Difference-Making Conditionals and the Relevant Ramsey Test. *The Review of Symbolic Logic*. doi: https://doi.org/10.1017/S1755020319000674

Rottman, B. M., & Hastie, R. (2014). Reasoning about causal relationships: Inferences on causal networks. *Psychological Bulletin, 140*(1), 109–139.

Sikorski, M., van Dongen, N. & Sprenger, J. (2019). *Causal Conditionals, Tendency Causal Claims and Statistical Relevance.* [Preprint] Retrieved from: http://philsci-archive.pitt.edu/16732/

Shipley, B. (2016). *Cause and Correlation in Biology*. Cambridge: Cambridge University Press.

Singmann, H., Bolker, B., Westfall, J. & Aust, F. (2020). *afex: Analysis of Factorial Experiments.* R package version 0.28-0. https://CRAN.R-project.org/package=afex

Skovgaard-Olsen, N. (2015). Ranking Theory and Conditional Reasoning. *Cognitive Science*. *40*(4), 848-880.

Skovgaard-Olsen, N. (2016). Motivating the Relevance Approach to Conditionals. *Mind and Language*, *31*(5), 555–579.

Skovgaard-Olsen, N., Singmann, H., & Klauer, K. C. (2016). The relevance effect and conditionals. *Cognition*, *150*, 26–36.

Skovgaard-Olsen, N., Singmann, H., & Klauer, K. C. (2017). Relevance and Reason Relations. *Cognitive Science*, *41*(S5), 1202–1215.

Skovgaard-Olsen, N., Collins, P., Krzyżanowska, K., Hahn, U., & Klauer, K. C. (2019). Cancellation, negation, and rejection. *Cognitive Psychology*, *108*, 42–71.

Skovgaard-Olsen, N., Kellen, D., Hahn, U., & Klauer, K. C. (2019). Norm conflicts and conditionals. *Psychological Review*, *126*(5), 611–633.

Skovgaard-Olsen, N., Kellen, D., Krahl, H., & Klauer, K. C. (2017). Relevance differently affects the truth, acceptability, and probability evaluations of “and”, “but”, “therefore”, and “if–then.” *Thinking and Reasoning*, *23*(4), 449–482.

Sloman, S. A. (2005). *Causal Models. How People Think about the World and its Alternatives*. Oxford: Oxford University Press.

Sloman, S. A. and Lagnado, D. A. (2005). Do We "do"? *Cognitve Science, 29*, 5-39.

Spohn, W. (2010). The Structural Model and the Ranking Theoretic Approach to Causation:

A Comparison. In Dechter, R., Geffner, H., Halpern, J. Y. (Eds.), *Heuristics, Probability and Causality. A Tribute to Judea Pearl* (pp. 493-508). San Mateo, CA: Kauffmann.

Spohn, W. (2012a). *The Laws of Beliefs*. Oxford: Oxford University Press.

Spohn, W. (2012b). Reversing 30 Years of Discussion: Why Causal Decision Theorists Should One-Box. *Synthese, 187*(1), 95–122.

Spohn, W. (2013). A ranking-theoretic approach to conditionals. *Cognitive Science, 37*(6), 1074–1106.

Stalnaker, R. C. (1968). A Theory of Conditionals. In: Rescher, N. (Eds.), *Studies in Logical Theory (pp. 98-112)*. Oxford: Basil Blackwell.

Stephan, S. & Waldmann, M. R. (2018). Preemption in singular causation judgments: A computational model. *Topics in Cognitive Science, 10,* 242-257.

Vance, J., & Oaksford, M. (2020). Explaining the implicit negations effect in conditional inference: Experience, probabilities, and contrast sets. *Journal of Experimental Psychology: General.* https://doi.org/10.1037/xge0000954

Vandenburgh, J. (2020). Conditional learning through causal models. *Synthese*. https://doi.org/10.1007/s11229-020-02891-x

VanderWeele, T. J. (2015). *Explantion in Causal Inference.* Oxford: Oxford University Press.

van Rooij, R., & Schulz, K. (2019). Conditionals, Causality and Conditional Probability. *Journal of Logic, Language and Information*, *28*(1), 55–71.

Vidal, M. (2017). A compositional semantics for 'even if' conditionals. *Logic and Logical Philosophy, 26,* 237-276.

Vidal, M., & Baratgin, J. (2017). A psychological study of unconnected conditionals. *Journal of Cognitive Psychology*, *29*(6), 769–781.

Waldmann, M. R. (1996). Knowledge-based causal induction. In D. R. Shanks, K. J. Holyoak, & D. L. Medin (Eds.), *The psychology of learning and motivation: Vol 34. Causal learning* (pp. 47–88). San Diego, CA: Academic Press.

Waldmann, M. R. (Eds.). (2017). *The Oxford Handbook of Causal Reasoning*. Oxford: Oxford University Press.

Waldmann, M. R., and Hagmayer, Y. (2005). Seeing versus Doing: Two Modes of Accessing Causal Knowledge. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*(2), 216-227.

Walton, D. (2004). *Relevance in Argumentation*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Wilson, D. and Sperber, D. (2004). Relevance Theory. In Horn, L.R. & Ward, G. (eds.), *The Handbook of Pragmatics*. Oxford: Blackwell, 607-632.

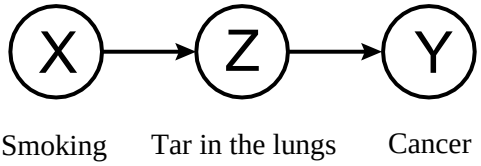
Appendix A: Bayes Nets and SEM

We here illustrate the difference between causal Bayes nets and structural equation modelling (SEM) in Pearl's (2009) theory of causal inference. While Pearl (1988) earlier argued that one could explain causal inferences solely in terms of causal Bayes nets, he later revised this account due to the need for structural equation models for counterfactual reasoning (Pearl, 2009; Pearl & Mackenzie, 2018).

On Pearl's (2009) current account, there are three irreducible layers of conceptual understanding of causal relations: 1) statistical associations for predictive inference (which can be computed by conditionalization, e.g. via Bayes nets), 2) predictions based on interventions (which are observed through manipulations in randomized, experimental studies),[23] and 3) counterfactual inferences (which can only be computed based on structural equation models of the data generating processes, as we show below).

[23] In addition, these interventions can now also be computed by applying Pearl's (2009) do-calculus to observational studies (see also Morgan & Winship, 2018).

Bayes nets encode a set of conditional independence statements to simplify the specification of a joint probability distribution over a set of causally relevant variables (Darwiche, 2009), such as the following:



Smoking Tar in the lungs Cancer

Figure A1. Bayes net representing a causal chain.

To illustrate their use in answering the queries above, the occurrence of the effect (e.g. cancer) can be predicted by conditionalizing on information about its possible causes (e.g. smoking), P(cancer|smoking). To evaluate the effect of an intervention (e.g. a hypothetical treatment designed to remove tar in the lungs), graph-surgery can be performed on the Bayes net. Graph surgery works by removing all incoming edges to the node intervened on, setting it to a given value (e.g. Z=0), and calculating the effects of the intervention on the descending nodes, P(cancer|*do*(tar=0)).

Finally, we can evaluate the counterfactual scenario in which we consider whether the the patient *would have been* cured, if the tar *had been* removed. But we need to make this evaluation while taking into account that the patient is in fact in a condition in which he has cancer and tar in his lungs. As a result, we need to be able to do both: 1) conditionalize on the factual information (cancer=1, tar=1) to update our distribution of the boundary conditions (U) representing the actual circumstances, and 2) perform graph surgery to calculate the effects of our counterfactual intervention. However, this latter step is not possible without structural equations representing the causal mechanisms underlying the causal diagram, which are shown below in Figure A2.

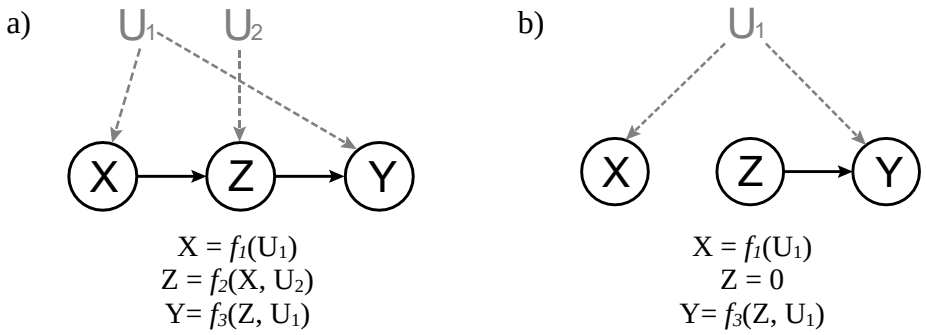


Figure A2. Left-side: structural equation model of the causal chain in Figure A1 with structural equations determining the values of endogenous variables X, Y, Z as a function of their parents and the exogenous variables, U_1 and U_2, representing the boundary conditions. Right-side: sub-model obtained by performing graph surgery on a) by replacing the equation for Z with Z = 0 and removing all edges to Z.

In this case, the boundary conditions might be unknown factors influencing both the amount of tar in the patient's lungs (U_2) and whether the patient smokes and has cancer (U_1). The structural equations in Figure A2 are used to update the distribution of the boundary conditions based on the available evidence, P(U | smoke=1, cancer=1, tar=1). This updated distribution *remains invariant* when considering the counterfactual scenario in which an intervention is introduced to set tar=0, through graph surgery to generate the submodel displayed as b) above. Finally, the counterfactual probability, $P(cancer=0_{tar=0} \mid cancer=1, tar=1)$, is calculated based on both the updated distribution of the boundary conditions and the submodel, where the graph surgery has been applied (Pearl, 2009, Ch. 7). Since Bayes nets lack structural equations representing the influence of the boundary conditions, Bayes nets cannot handle cases, where we both update based on the evidence (cancer=1, tar=1) and consider *what would have* happened if tar *had been* 0 under the actual circumstances.

Formally, this double evaluation of 1) an update by factual information (cancer=1, tar=1) concerning the actual world and 2) computation of probabilities in counterfactual scenarios (cancer=0, tar=0) would give rise to inconsistency, if represented by standard probability theory via conditionalization alone. The use of structural models illustrated in Figure 2A prevents this by separating the update that is kept invariant between the models a) and b), and computating the counterfactual update in the submodel b) only. To represent this

type of computational query, Pearl introduces the notation $P(Y_{x'} = \text{false} \mid X = \text{true}, Y = \text{true})$. In words: under the assumption that both events actually occurred ($P(\cdot \mid X = \text{true}, Y = \text{true})$),[24] what is the probability that Y would not have occurred had X not occurred ($P(Y_{x'} = \text{false} \mid \cdot)$).

Sometimes the term 'Structural Causal Model' (SCM) is used by Pearl to emphasize the integration of SEM as a statistical tool with causal graphs, a counterfactual semantics, and an explicit causal interpretation of strutural equations. Recent books on SEM have integrated many of these developments (see e.g. Kline, 2016; Shipley, 2016). We provide further details on structual equation models, when we apply them as a statistical tool in Experiment 3.

Appendix B: Simulation Analysis, Causal Power

In Appendix B, we consider the option of adopting a causal power account while dropping van Rooij and Schulz's (2019) auxillary assumption that participants' tendency to ignore alternative causes make them evaluate P(if A, then C) as P(C|A). Instead, the causal power account of the acceptability of indicative conditionals could be strengthened by the observation that the equation in Cheng (1997) requires causal power and P(C|A) to be highly correlated for generative causes. This observation might in turn account for the positive association between P(if A, then C) and P(C|A). To examine exactly how strongly P(C|A) and P(if A, then C) would be associated on a pure causal power account, a simulation analysis was carried out with 488422 probability distributions generated through gridsearch (see Table B1, upper part):

Table B1. Simulation Analysis, Pure Casual Power Account

	r(Y, P(C|A))	**r(Y, P(C|¬A))**	**m1: (Y ~ P(C|A))**	**m2: (Y ~ P(C|A) + P(C|¬A))**
		Simulation		
Y = power	$r_{Y,P(C\|A)} = .82$	$r_{Y,P(C\|\text{-}A)} = 0.03$	$\beta1 = .82$,	$\beta1 = 1.08, \beta2 = -.52$
	$r_{Y,P(C\|A).P(C\|\text{-}A)} = .94$	$r_{Y,P(C\|\text{-}A).P(C\|A)} = -.79$	$R^2 = .68$	$R^2 = .88$
Y = ΔP	$r_{Y,P(C\|A)} = .5$	$r_{Y,P(C\|\text{-}A)} = -.5$	$\beta1 = .5$,	$\beta1 = 1.0, \beta2 = -1.0$
	$r_{Y,P(C\|A).P(C\|\text{-}A)} = 1.0$	$r_{Y,P(C\|\text{-}A).P(C\|A)} = -1.0$	$R^2 = .25$	$R^2 = 1$
		Experiment 1		
Y = If	$r_{Y,P(C\|A).P(C\|\text{-}A)} = .72$	$r_{Y,P(C\|\text{-}A).P(C\|A)} = .04$	$\beta1 = .75$	$\beta1 = .74, \beta2 = .03$
Y = power	$r_{Y,P(C\|A).P(C\|\text{-}A)} = .69$	$r_{Y,P(C\|\text{-}A).P(C\|A)} = -.42$	$\beta1 = .61$	$\beta1 = .74, \beta2 = -.37$

[24] The dot, $\cdot$, is here used as a placeholder for an event, proposition, or random variable.

Note. The comparison is based on Positive Relevance conditions only. Upper half: correlation and least square regression analysis of simulated data based on 488422 probability distributions, which were generated meeting the criterion of Positive Relevance. Lower half: reanalysis of the Positive Relevance condition of Experiment 1 based on mixed regression models.

As the simulation shows, it is required that a causal power construct not only is strongly positively associated with P(C|A), $\beta 1 = 1.08$, in a regression analysis, but also negatively associated with P(C|¬A), $\beta 2 = -.52$.

In Over et al. (2007), it was assumed that on a causal analysis, it would be required that the negative assocaition of P(C|¬A) with P(if A, then C) would be of the same magnitude as the positive association of P(C|A). However, as the simulation analysis shows, this constraint only holds for ΔP. In contrast, on a causal power account, the absolute magnitude of the positive association of P(C|A) is twice that of the negative association with P(C|¬A). Nevertheless, in previous studies—like Evans et al. (2007) and Over et al. (2007)—it was found that although weak, negative associations between P(C|¬A) and P(if A, then C) did occur, they were of a much smaller magnitude than the ones shown above.

In the lower part of Table B1, a reanalysis of parts of the data from Experiment 1 was carried out with the type of mixed regression model reported in Table 5. This type of model also contains a random intercept controlling for differences between scenarios, while estimating fixed, mean effects. The required negative association of P(C|¬A) with P(if A, then C) was not obtained for this subset of the data. P(C|¬A) was, however, negatively associated with causal power. Thus, like the model comparison in Table 5, this reanalysis did not turn out favorably for a causal power account of P(if A, then C).